BAYESIAN ANALYSIS FOR LARGE SPATIAL DATA

A Dissertation

by

JINCHEOL PARK

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2012

Major Subject: Statistics

BAYESIAN ANALYSIS FOR LARGE SPATIAL DATA

A Dissertation

by

JINCHEOL PARK

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Faming Liang
Committee Members,	Marc Genton
	Michael Sherman
	Jianxin Zhou
Head of Department,	Simon J. Sheather

August 2012

Major Subject: Statistics

ABSTRACT

Bayesian Analysis for Large Spatial Data. (August 2012)
Jincheol Park, MSc., Statistics, University of Ottawa, 2000, Canada;
BSc., Statistics, Seoul National University, 1997, South Korea
Chair of Advisory Committee: Faming Liang

The Gaussian geostatistical model has been widely used in Bayesian modeling of spatial data. A core difficulty for this model is at inverting the $n \times n$ covariance matrix, where n is a sample size. The computational complexity of matrix inversion increases as $O(n^3)$. This difficulty is involved in almost all statistical inferences approaches of the model, such as Kriging and Bayesian modeling. In Bayesian inference, the inverse of covariance matrix needs to be evaluated at each iteration in posterior simulations, so Bayesian approach is infeasible for large sample size n due to the current computational power limit.

In this dissertation, we propose two approaches to address this computational issue, namely, the auxiliary lattice model (ALM) approach and the Bayesian site selection (BSS) approach. The key feature of ALM is to introduce a latent regular lattice which links Gaussian Markov Random Field (GMRF) with Gaussian Field (GF) of the observations. The GMRF on the auxiliary lattice represents an approximation to the Gaussian process. The distinctive feature of ALM from other approximations lies in that ALM avoids completely the problem of the matrix inversion by using analytical likelihood of GMRF. The computational complexity of ALM is rather attractive, which increase linearly with sample size.

The second approach, Bayesian site selection (BSS), attempts to reduce the dimension of data through a smart selection of a representative subset of the observations. The BSS method first split the observations into two parts, the observations

near the target prediction sites (part I) and their remaining (part II). Then, by treating the observations in part I as response variable and those in part II as explanatory variables, BSS forms a regression model which relates all observations through a conditional likelihood derived from the original model. The dimension of the data can then be reduced by applying a stochastic variable selection procedure to the regression model, which selects only a subset of the part II data as explanatory data. BSS can provide us more understanding to the underlying true Gaussian process, as it directly works on the original process without any approximations involved.

The practical performance of ALM and BSS will be illustrated with simulated data and real data sets.

ACKNOWLEDGMENTS

First of all, I give thanks to the Lord, for by the grace of God I am what I am. It has been the Lord's grace that leads and strengthens me to finish my study.

I would like to express my great gratitude to my supervisor, Professor Faming Liang, for his valuable inspiration and guidance.

I also acknowledge the grateful encouragement and consistent prayers from my families in Canada and Korea.

Finally, I am truly grateful to my wife, Jee Young Yang, and my son, Joseph Hayoung, for being beside me with precious prayers and supports.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION*

Geostatistics is a branch of spatial statistics which deals with the data obtained by sampling a spatially continuous process $\{X(s)\}$, $s \in \mathbb{R}^2$, at a discrete set of locations $\{s_i, i = 1, \cdots, n\}$ in a spatial region of interest $A \subset \mathbb{R}^2$. Consider a Gaussian geostatistical model,

$$\begin{aligned} Y(s_i) &= \nu(s_i) + X(s_i) + \varepsilon_i, \\ \varepsilon_i &\overset{iid}{\sim} N\left(0, \tau^2\right), \end{aligned} \tag{1.1}$$

where $\{Y(s_i)\}$ denotes our observations at locations $s_1, \cdots, s_n$, $\{\nu(s_i)\}$ denotes the mean of $\{Y(s_i)\}$, $\{X(s_i)\}$ denotes a spatial Gaussian process with $E\{X(s_i)\} = 0$, $Var\{X(s_i)\} = \sigma^2$, and $Corr\{X(s_i), X(s_j)\} = \rho(\|s_i - s_j\|)$ for an appropriate correlation function with Euclidean distance $\|\cdot\|$, and τ^2 is called the nugget variance in this context. The correlation function is chosen from some parametric families, such as the Matérn, powered exponential or spherical (Cressie, 1993). Under model (1.1), $\{Y(s)\}$ follows a multivariate Gaussian distribution,

$$Y(\boldsymbol{s}) \sim MVN(\boldsymbol{\nu}, V), \tag{1.2}$$

where $\boldsymbol{\nu} = \{\nu(s_1), \ldots, \nu(s_n)\}^T$, $V = \sigma^2\Sigma + \tau^2 I$, and I is the $n \times n$ identity matrix and Σ is an $n \times n$ matrix with the $(i, j)^{th}$ element being defined by $\rho(\|s_i - s_j\|)$. Model (1.1) is perhaps the most popular model in geostatistics. It can be easily extended to

---

This dissertation follows the style of *Journal of Computational and Graphical Statistics*.

*Reprinted with permission from "Bayesian Analysis of Geostatistical Models with an Auxiliary Lattice" by Park, J. and Liang, F., 2012, *Journal of Computational and Graphical Statistics*, Copyright by Talyor & Francis.

the regression setting with the mean $\{\nu(s)\}$ being replaced by

$$\nu(s_i) = \xi_0 + \sum_{j=1}^{p} \xi_j c_j(s_i), \tag{1.3}$$

where $c_j(\cdot)$ denotes the j^{th} explanatory variable, and ξ_j denotes the corresponding regression coefficient. Evaluation of the likelihood function of model (1.1) (Diggle *et al.*, 1998) involves inverting an $n \times n$ matrix—the covariance matrix V. It is known that the computational complexity of matrix inversion increases as $O(n^3)$. When n is large, this is infeasible due to the current computational power limit.

A variety of methods for tackling this obstacle have been proposed in the literature. These methods can be roughly grouped into three categories, lower dimensional space approximation, likelihood approximation, and sparse matrix-based approximation.

The methods in the first category seek to approximate the spatial process $\{X(s)\}$ by a lower dimensional space process $\{\tilde{X}(s)\}$ with the use of smoothing techniques, such as kernel convolutions, moving averages, low rank splines, or basis functions, see e.g., Wikle and Cressie (1999), Lin *et al.* (2000), Kammann and Wand (2003), Paciorek (2007), Banerjee *et al.* (2008), and Finley *et al.* (2009). The method of continuous global surfaces (Billings *et al.*, 2002), which provides a framework of interpolation and smoothing for geophysical data, also falls into this category. Although these methods can reduce the computational burden to some extent, it cannot completely avoid a matrix inversion. For a large dataset, the dimension of the approximation process $\{\tilde{X}(s)\}$ can still be very high, rendering inapplicability of these methods.

The methods in the second category seek to approximate the likelihood function in spectral domain (see e.g., Fuentes, 2007) or by a product of conditional densities

(see, e.g., Vecchia, 1988; Jones and Zhang, 1997; and Stein *et al.*, 2004). Concerns with these methods include adequacy of the likelihood approximation and some implementation issues. Expertise is required for selecting an appropriate spectral density estimate or a sequence of conditional densities. In addition, the spectral density methods are best suited to stationary covariance functions.

The methods in the third category are to approximate the spatial process $\{X(s)\}$ by a manageable process for which the covariance matrix is sparse. Examples of such approximations include covariance tapering and Markov random field approximations.

In covariance tapering method (see e.g., Furrer *et al.*, 2006; Kaufman *et al.*, 2008; Furrer and Bengtsson, 2007), elements of the covariance matrix corresponding to spatially distant pairs of observations are set to zero in a way to retain positive definiteness property of the resulting matrix.

The Markov random field approximation method (see e.g., see e.g., Rue and Tjelmeland, 2002; Rue and Held, 2005), as suggested by its names, is to approximate the spatial process by a Markov random field whose covariance matrix is known to be sparse. This method is first proposed for regularly spaced data. For irregularly spaced data, Hartman and Hössjer (2008) suggested to approximate the spatial process by a Markov random field on a lattice and then interpolate the irregularly spaced data based on the estimates at the grid points of the lattice. Besag and Mondal (2005) also discussed the possibility of extending the de Wijs process, a Gaussian Markov process, approximation to irregularly spaced data.

Recently, Rue et al. (2009) suggested the integrated nested Laplace approximation (INLA) method for approximate Bayesian inference of latent Gaussian models. Lindgren *et al.* (2010) applied INLA to the Gaussian field by representing it as a Gaussian Markov random field (GMRF) through solving a stochastic partial differential equation (SPDE). However, as pointed out in Lindgren *et al.* (2010), one

drawback with SPDE approach is that there exist explicit representations of GMRFs only for those Gaussian fields having a Matern Covariance stducture at certain interger smoothness. For example, Gaussian fields with exponential covariance structure is not able to be represented explicitly by GMRFs through the SPDE approach. (See e.g., Sun *et al.*, 2012 for a detailed review).

In the Chapter II, we propose auxiliary lattice model (ALM) which is a hierarchical model for large irregularly spaced data by introducing an auxiliary regular lattice to the space of observations. We define a Gaussian Markov random field (GMRF) on the auxiliary lattice, which represents an approximation to the process $\{X(s)\}$. This is motivated by the observation of Rue and Tjelmeland (2002): The GMRF with small neighborhoods can approximate Gaussian random fields surprisingly well even with long correlation lengths. Conditioned on the GMRF, we model $\{X(s)\}$ under a regression setting such that $[X(s_i)|Z]$'s are mutually independent, where Z denotes the GMRF defined on the auxiliary lattice, and $[\cdot|Z]$ denotes the conditional distribution of $X(s_i)$. In spirit, ALM is similar to the method proposed by Hartman and Hössjer (2008). However, ALM completely avoids the problem of matrix inversion by using analytical results of GMRFs, and thus ALM can have a better scalability than Hartman and Hössjer's method. Note that the computational complexity of the sparse matrix-based methods is $O(n^*)$ and is hard to reach $O(n)$, where n^* denotes the number of nonzero elements in the covariance matrix. However, as discussed in Sections D and E in Chapter II, the computational complexity of our method is $O(n)$, which implies that ALM can be applied to very large datasets with reasonable CPU times. In addition, ALM works under the Bayesian framework, so it can have a better measurement for the uncertainty of parameter estimates and prediction. The numerical results show that ALM can approximate Gaussian random fields very well in terms of predictions, even for those with long correlation lengths. For real data

examples, ALM can generally outperform the conventional Gaussian random field models in both prediction errors and CPU times.

A problem of general interest in spatial statistics is to predict unobserved values of $\{Y(s_i^p)\}$ at a set of locations $\boldsymbol{s}^p = \{s_1^p, \ldots, s_{n_p}^p\}$. In Chapter III, we propose a prediction-oriented Bayesian site selection (BSS) method which, while reducing the dimension of data, attempts to avoid the shortcomings of the dependence truncation, lower-dimensional process approximation, and likelihood approximation methods. The BSS method first split the observations into two parts, the observations near the prediction sites (part I) and their remaining (part II). Then, by treating the observations in part I as response variable and those in part II as explanatory variables, BSS forms a regression model which relates all observations $\{Y(s_i)\}$ through a conditional likelihood derived from the original model (1.1). The dimension of the data can then be reduced by applying a stochastic variable selection procedure to the regression model, which selects only a subset of the part II data as explanatory variables. The selected explanatory variables together with the response data thus form the basis of observations for inference of model (1.1) and prediction of unobserved values. Compared to the dependence truncation methods, BSS is able to catch the long range dependence through selection of appropriate explanatory variables. Compared to the lower-dimensional process and likelihood approximation methods, BSS can provide us more understanding to the underlying true Gaussian process, as it directly works on the original process without any approximations involved.

CHAPTER II

BAYESIAN AUXILIARY LATTICE MODEL*

A. Auxiliary Gaussian Markov Random Field

Consider a spatial dataset $\{Y(s_i)\}$ observed on a set of locations $s_1, \ldots, s_n$. To model the observations, we introduce an $M \times N$ auxiliary square lattice $W = \{(k,l) : k = 1, \cdots, M, l = 1, \cdots, N\}$ to the space of observations, as illustrated by Figure 1, which covers the region of interest. Let $s^w_{kl} = (s^w_k, s^w_l) \in \mathbb{R}^2$ denote the geographical location of the grid point (k,l) of W. Then we define a zero-mean GMRF, which is denoted by $Z = \{Z_{kl}, (k,l) \in W\}$ throughout this dissertation, on the auxiliary lattice. Let $\boldsymbol{z}$ denote a realization of Z. Its log-likelihood function is given by

$$\log f(\boldsymbol{z}|\boldsymbol{\beta}, \sigma^2) = -\frac{MN}{2}\log(2\pi) - \frac{MN}{2}\log(\sigma^2) + \frac{1}{2}\log|Q(\boldsymbol{\beta})| - \frac{1}{2\sigma^2}\vec{\boldsymbol{z}}^T Q(\boldsymbol{\beta})\vec{\boldsymbol{z}},$$

where $Q(\boldsymbol{\beta})$ is the potential matrix and $\boldsymbol{\beta}$ is a vector containing the interaction parameters of the GMRF, and $\vec{\boldsymbol{z}}$ denotes a prolonged vector of $\boldsymbol{z}$, which is arranged by rows, with z_{kl} being its $((k-1) \times N + l)$-th element; that is,

$$\vec{\boldsymbol{z}} = (z_{11}, z_{12}, \ldots, z_{1N}, z_{21}, \ldots, z_{2N}, \ldots, z_{M1}, \ldots, z_{MN})^T.$$

The neighborhood structure of the GMRF is illustrated by (2.1), where the left part shows a second-order isotropic neighborhood structure and the right part shows the coefficients of pair interactions.

---

*Reprinted with permission from "Bayesian Analysis of Geostatistical Models with an Auxiliary Lattice" by Park, J. and Liang, F., 2012, *Journal of Computational and Graphical Statistics*, Copyright by Talyor & Francis.

$$\begin{vmatrix} \times & \times & \times & \times & \times \\ \times & 2 & 1 & 2 & \times \\ \times & 1 & o & 1 & \times \\ \times & 2 & 1 & 2 & \times \\ \times & \times & \times & \times & \times \end{vmatrix} \qquad \begin{vmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_d & \beta_v & \beta_d & 0 \\ 0 & \beta_h & 1 & \beta_h & 0 \\ 0 & \beta_d & \beta_v & \beta_d & 0 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix} \tag{2.1}$$

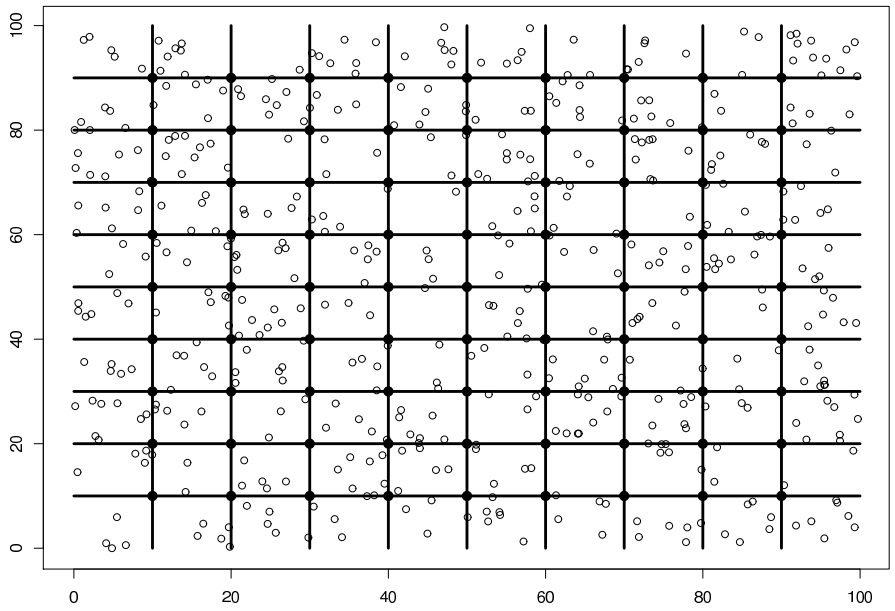

Fig. 1. Illustration of the auxiliary lattice: The black points denote the sites on which the auxiliary GMRF Z is defined.

If we assume Z has a second-order neighborhood structure and a free boundary condition, then $\boldsymbol{\beta}$ consists of three interaction parameters $(\beta_h, \beta_v, \beta_d)$ and the

potential matrix $Q(\boldsymbol{\beta})$ can be written as

$$Q(\boldsymbol{\beta}) = \begin{bmatrix} A & B & 0 & \cdots & 0 \\ B & A & B & & \\ 0 & B & A & B & \vdots \\ \vdots & \vdots & \ddots & \ddots & B \\ 0 & \cdots & & B & A \end{bmatrix} \tag{2.2}$$
$$= I_M \otimes A + S_M \otimes B,$$

where

$$A = \begin{bmatrix} 1 & -\beta_h & 0 & \cdots & 0 \\ -\beta_h & 1 & -\beta_h & 0 & \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & -\beta_h & 1 & -\beta_h \\ & & & -\beta_h & 1 \end{bmatrix} = I_N - \beta_h S_N,$$

$$B = \begin{bmatrix} -\beta_v & -\beta_d & 0 & \cdots & 0 \\ -\beta_d & -\beta_v & -\beta_d & 0 & \\ 0 & & \ddots & \ddots & \vdots \\ \vdots & \vdots & & -\beta_v & -\beta_d \\ & & & -\beta_d & -\beta_v \end{bmatrix} = -\beta_v I_N - \beta_d S_N,$$

with I_k being a $k \times k$ identity matrix and S_k being a $k \times k$ matrix such that

$$S_k = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & & \\ 0 & 1 & & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}_{k \times k}.$$

The eigenvalues of the potential matrix $Q(\boldsymbol{\beta})$ are given by

$$\lambda_{kl}(Q(\boldsymbol{\beta})) = 1 - 2\beta_v \cos\frac{k\pi}{M+1} - 2\beta_h \cos\frac{l\pi}{N+1} - 4\beta_d \cos\frac{k\pi}{M+1}\cos\frac{l\pi}{N+1},$$

for $1 \le k \le M$ and $1 \le l \le N$. See Balram and Moura (1993) for the proof of this result. Then the likelihood function of Z can be analytically expressed as

$$\begin{aligned}\log f(\boldsymbol{z}|\boldsymbol{\beta},\sigma^2) = &-\frac{MN}{2}\log 2\pi - \frac{MN}{2}\log\sigma^2 - \frac{MN}{2\sigma^2}\left(T_t - 2\beta_h T_h - 2\beta_v T_v - 2\beta_d T_d\right) \\ &+ \frac{1}{2}\sum_{k=1}^{M}\sum_{l=1}^{N}\log\left(1 - 2\beta_v\cos\frac{k\pi}{M+1} - 2\beta_h\cos\frac{l\pi}{N+1} - 4\beta_d\cos\frac{k\pi}{M+1}\cos\frac{l\pi}{N+1}\right),\end{aligned} \tag{2.3}$$

where

$$T_t = \frac{1}{MN}\sum_{k=1}^{M}\sum_{l=1}^{N} z_{kl}^2, \quad T_h = \frac{1}{MN}\sum_{k=1}^{M}\sum_{l=1}^{N-1} z_{kl}z_{k(l+1)}, \quad T_v = \frac{1}{MN}\sum_{k=1}^{M-1}\sum_{l=1}^{N} z_{kl}z_{(k+1)l},$$

$$T_d = \frac{1}{MN}\left\{\sum_{k=1}^{M-1}\sum_{l=1}^{N-1} z_{kl}z_{(k+1)(l+1)} + \sum_{k=1}^{M-1}\sum_{l=2}^{N} z_{kl}z_{(k+1)(l-1)}\right\}.$$

For a higher-order neighborhood structure, the eigenvalues of the potential matrix $Q(\boldsymbol{\beta})$ can be calculated using the technique of discrete Fourier transformation (see e.g., Rue and Held, 2005), so the analytical form of the likelihood function is also available. In this case, a torus boundary condition may be assumed for the lattice. This boundary condition is reasonable for a large lattice. For an illustration purpose, we consider only the second-order neighborhood structure in this Chapter.

B. A Hierarchical Geostatistical Model

Conditioned on a GMRF Z, we assume that $X(s_i)$'s are mutually independent; that is, the density function can be factorized as

$$f\{x(s_1),\ldots,x(s_n)|\boldsymbol{z}\} = f\{x(s_1)|\boldsymbol{z}\}\cdots f\{x(s_n)|\boldsymbol{z}\}. \tag{2.4}$$

In addition, we assume that $\{X(s_i), \vec{Z}\}$ is distributed as a multivariate Gaussian distribution with mean zero and the covariance matrix given by

$$\Sigma_i = \sigma^2 \begin{bmatrix} 1 & \boldsymbol{r}_i^T \\ \boldsymbol{r}_i & Q(\boldsymbol{\beta})^{-1} \end{bmatrix}, \tag{2.5}$$

where $\boldsymbol{r}_i = Corr\{X(s_i), \vec{Z}\}$ with its $\{(k-1) \times N + l\}$-th element being defined as $Corr\{X(s_i), Z_{kl}\}$. Theorem 1 shows that if $1 - \boldsymbol{r}_i^T Q(\boldsymbol{\beta}) \boldsymbol{r}_i > 0$, then Σ_i is positive definite. The proof of this theorem can be found in the Appendix.

Theorem 1 *Suppose that $Q(\boldsymbol{\beta})$ is a positive definite matrix and $1 - \boldsymbol{r}_i^T Q(\boldsymbol{\beta}) \boldsymbol{r}_i > 0$ for all $i = 1, \cdots, n$. Then Σ_i defined in (2.5) is also a positive definite matrix for all $i = 1, \cdots, n$.*

Following the simple kriging theory, we have

$$X(s_i)|Z \sim N\left\{\boldsymbol{r}_i^T Q(\boldsymbol{\beta}) \vec{Z}, \sigma^2(1 - \boldsymbol{r}_i^T Q(\boldsymbol{\beta}) \boldsymbol{r}_i)\right\}, \tag{2.6}$$

which is equivalent to assuming a regression relationship between X and Z with the random errors being independently and normally distributed.

Note that $X(s_i)$ is not necessarily correlated with all Z_{kl}'s. In Section D, we consider a simplified version of this model, for which $X(s_i)$ is only correlated with a subset of Z. In this dissertation, we suggest to choose the correlation function between X and Z in the spherical family; that is,

$$\text{Corr}\{X(s_i), Z_{kl}\} = \begin{cases} 1 - \frac{3}{2}\frac{h_{i(kl)}}{\phi} + \frac{1}{2}\left(\frac{h_{i(kl)}}{\phi}\right)^3, & 0 \le h_{i(kl)} \le \phi, \\ 0, & h_{i(kl)} > \phi, \end{cases} \tag{2.7}$$

where ϕ is a parameter, and $h_{i(kl)} = \|s_i - s_{kl}^w\|$ is the Euclidean distance between the sites of $X(s_i)$ and Z_{kl}. Compared to the corelation functions in the Matérn and

powered exponential families, (2.7) reduces computational complexity of our model due to its tail truncation. However, a compact support of the correlation function is not an essential requirement for our model.

Let $\boldsymbol{\theta}$ denote the parameter vector of our model, which includes τ^2, σ^2, the regression coefficients $\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_p)$ in (1.3), the correlation range parameter ϕ in (2.7), and the interaction parameters $\boldsymbol{\beta}$ of the auxiliary GMRF. With the assumptions (2.4)–(2.6), the likelihood of our new model can be written as

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta}) &= \int f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) f(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\theta}) d\boldsymbol{x} \\ &= \prod_{i=1}^{n} \int f\{y(s_i)|x(s_i),\boldsymbol{\theta}\} f\{x(s_i)|\boldsymbol{z},\boldsymbol{\theta}\} dx(s_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(\tau^2+\sigma_i^2)}} \exp\left[-\frac{1}{2\left(\tau^2+\sigma_i^2\right)} \left\{ y(s_i) - \nu(s_i) - \mu_i \right\}^2 \right], \end{aligned} \tag{2.8}$$

where $\sigma_i^2 = \sigma^2(1 - \boldsymbol{r}_i^T Q(\boldsymbol{\beta}) \boldsymbol{r}_i)$, $\mu_i = \boldsymbol{r}_i^T Q(\boldsymbol{\beta}) \vec{\boldsymbol{z}}$. For convenience, we will, henceforth, call the new model the *auxiliary lattice Gaussian* model and denote it by M_{AL} in a shorthand notation. Correspondingly, we will denote the Gaussian model (1.1) by M_G.

To see the advantage of the model M_{AL}, we first obtain $f(\boldsymbol{y}|\boldsymbol{\theta})$ by integrating $\boldsymbol{z}$ out from (2.8),

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}) &= \int f(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta}) f(\boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{z} \\ &= |2\pi\Sigma_y|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}-\boldsymbol{\nu}) \Sigma_y^{-1} (\boldsymbol{y}-\boldsymbol{\nu}) \right\}, \end{aligned} \tag{2.9}$$

where the covariance matrix Σ_y is given by

$$\begin{aligned}\Sigma_y &= \begin{bmatrix} \tau^2+\sigma^2 & \boldsymbol{r}_1^T Q(\boldsymbol{\beta})\boldsymbol{r}_2 & \cdots & \cdots & \boldsymbol{r}_1^T Q(\boldsymbol{\beta})\boldsymbol{r}_n \\ \boldsymbol{r}_2^T Q(\boldsymbol{\beta})\boldsymbol{r}_1 & \tau^2+\sigma^2 & \boldsymbol{r}_2^T Q(\boldsymbol{\beta})\boldsymbol{r}_3 & \cdots & \boldsymbol{r}_2^T Q(\boldsymbol{\beta})\boldsymbol{r}_n \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \boldsymbol{r}_n^T Q(\boldsymbol{\beta})\boldsymbol{r}_1 & \cdots & \cdots & \boldsymbol{r}_n^T Q(\boldsymbol{\beta})\boldsymbol{r}_{n-1} & \tau^2+\sigma^2 \end{bmatrix} \\ &= Diag[\tau^2+\sigma_1^2,\ldots,\tau^2+\sigma_n^2] + R^T Q(\boldsymbol{\beta})R \triangleq D + R^T Q(\boldsymbol{\beta})R \end{aligned} \tag{2.10}$$

and $R = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_n]$ is a matrix formed by r_i's. Following from (2.9) and (2.10), we can rewrite the auxiliary lattice model as

$$\begin{aligned} Y(s_i) &= \nu(s_i) + \boldsymbol{r}_i^T Q(\boldsymbol{\beta})\vec{Z} + \epsilon_i, \\ \epsilon_i &\sim N(0, \sigma_i^2 + \tau^2), \end{aligned} \tag{2.11}$$

where ϵ_i's are independent, and the term $\boldsymbol{r}_i^T Q(\boldsymbol{\beta})\vec{Z}$ corresponds to the simple kriging prediction at s_i based on the auxiliary GMRF Z. Applying the Woodbury identity, we have

$$\Sigma_y^{-1} = \{D + R^T Q(\boldsymbol{\beta})R\}^{-1} = D^{-1} - D^{-1}R^T\{Q^{-1}(\boldsymbol{\beta}) + RD^{-1}R^T\}^{-1}RD^{-1}. \tag{2.12}$$

Hence, inverting Σ_y^{-1} can be reduced to invert a $(MN \times MN)$-matrix $\{Q(\boldsymbol{\beta})^{-1} + RD^{-1}R^T\}$, if MN is smaller than n. It is true that the lower dimensional process approximation methods can lead to a significant reduction in computation, however, for large datasets, it is often the case that inverting the dimension-reduced matrix is still a cumbersome task. As reviewed in the Introduction, this is also the drawback that most dimension reduction methods suffer from.

In this dissertation, we consider a data augmentation scheme (Tanner and Wong, 1987) for sampling from the posterior of the auxiliary lattice model by treating Z as "missing" values. This converts the matrix inversion problem to a sampling problem:

(i) Sampling $\boldsymbol{z}$ conditional on $\boldsymbol{y}$ and $\boldsymbol{\theta}$;

(ii) Sampling $\boldsymbol{\theta}$ conditional on $\boldsymbol{y}$ and $\boldsymbol{z}$.

The details are given in Section C. The unique benefit of this sampling scheme is not in dimension reduction but in that it completely avoids matrix inversions.

C. Bayesian Analysis for Model M_{AL}

In this section, we consider Bayesian analysis for the model M_{AL} with the second-order neighborhood structure. Firstly, we a specify a noninformative prior for $\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_p)$ which are defined in (1.3); that is,

$$\pi(\boldsymbol{\xi}) \propto 1. \tag{2.13}$$

Let v_s^2 denote the sample variance of the data. Since $v_s^2 \geq \sigma^2 + \tau^2$ is generally true for a reasonably large dataset and it is generally believed that the nugget variance is smaller than the variance of the latent Gaussian process, it is reasonable to assume the following priors for σ^2 and τ^2:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} I(L_{\sigma^2} \leq \sigma^2 \leq U_{\sigma^2}), \quad \pi(\tau^2|\sigma^2) \propto \frac{1}{\tau^2} I(L_{\tau^2} \leq \tau^2 \leq \sigma^2), \tag{2.14}$$

for some positive numbers L_{σ^2}, U_{σ^2} and L_{τ^2}. For example, we may set $L_{\sigma^2} = 0.01 v_s^2$, $U_{\sigma^2} = 2v_s^2$ and $L_{\tau^2} = 0.001 v_s^2$. For the parameters $\boldsymbol{\beta}$ and ϕ, we assume

$$\pi(\boldsymbol{\beta}) \propto I(|\beta_h| + |\beta_v| + 2|\beta_d| < 0.5), \quad \pi(\phi|\boldsymbol{\beta}) \propto \prod_{i=1}^{n} I(1 - \boldsymbol{r}_i^T Q(\boldsymbol{\beta}) \boldsymbol{r}_i > 0) I(0 < \phi < U_\phi), \tag{2.15}$$

for some $U_\phi > 0$. In simulations, U_ϕ can be set to a large number, say 10^{10}. As a practical matter, this is equivalent to set $U_\phi = \infty$. Note that the prior of $\boldsymbol{\beta}$ ensures the stationarity of the GMRF as shown by Balram and Moura (1993), and the prior

of ϕ ensures the positive-definite covariance matrix of Σ_i as shown in Theorem 1.

With the above priors, the posterior of our model can be written as

$$f(\sigma^2, \tau^2, \phi, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{z}|\boldsymbol{y}) \propto \pi(\boldsymbol{\xi})\pi(\sigma^2)\pi(\tau^2|\sigma^2)\pi(\boldsymbol{\beta})\pi(\phi|\boldsymbol{\beta})f(\boldsymbol{z}|\sigma^2, \boldsymbol{\beta})f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}), \quad (2.16)$$

where $f(\boldsymbol{z}|\sigma^2, \boldsymbol{\beta})$ and $f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})$ are given in (2.3) and (2.8), respectively. It is easy to show that the marginal posterior of $\boldsymbol{\xi}$ is normal and thus proper. Since other priors are all proper, the joint posterior (2.16) is proper.

Let $\boldsymbol{s}^p = \{s_1^p, \cdots, s_{n_p}^p\}$ denote a set of locations to predict on. Then $Y(\boldsymbol{s}_p)$ can be predicted by its conditional mean $E\{Y(\boldsymbol{s}_p)|Y(\boldsymbol{s})\}$, which is given by

$$E\{Y(\boldsymbol{s}_p)|Y(\boldsymbol{s})\} = \int\int E\{Y(\boldsymbol{s}_p)|Y(\boldsymbol{s}), \boldsymbol{\theta}, \boldsymbol{z}\} f\{\boldsymbol{\theta}, \boldsymbol{z}|Y(\boldsymbol{s})\} d\boldsymbol{\theta} d\boldsymbol{z}.$$

It follows from (2.8) that

$$E\{Y(s_i^p)|Y(\boldsymbol{s}), \boldsymbol{\theta}, \boldsymbol{z}\} = \nu(s_i^p) + r_{p,i}^T Q(\boldsymbol{\beta})\vec{\boldsymbol{z}}, \quad i = 1, \ldots, n_p,$$

where $r_{p,i}^T = Corr\{X(s_i^p), \vec{Z}\}$. Hence, the prediction $E\{Y(\boldsymbol{s}_p)|Y(\boldsymbol{s})\}$ can be calculated using MCMC samples of $\boldsymbol{\theta}$ and Z simulated from (2.16).

To simulate samples from $f\{\boldsymbol{\theta}, \boldsymbol{z}|Y(\boldsymbol{s})\}$, we implement the data augmentation scheme in a manner of the Metropolis-within-Gibbs sampler (Müller, 1991) as follows. To facilitate sampling, the parameters $\boldsymbol{\beta}, \sigma^2, \phi, \tau^2, \boldsymbol{\xi}$ are reparameterized and then grouped into three subgroups: $\theta_1 = \{\boldsymbol{\beta}, \log(\sigma^2)\}$, $\theta_2 = \log(\phi)$ and $\theta_3 = \{\log(\tau^2), \boldsymbol{\xi}\}$. Let $\theta_1^{(t)}$, $\theta_2^{(t)}$, $\theta_3^{(t)}$ and $\boldsymbol{z}^{(t)}$ denote, respectively, the samples of θ_1, θ_2, θ_3 and Z drawn at iteration t. For notational simplicity, in what follows we depress the subscript t and denote by θ_1^*, θ_2^*, θ_3^* and $\boldsymbol{z}^*$ the samples of θ_1, θ_2, θ_3 and $\boldsymbol{z}$ drawn at iteration $t+1$.

(1) Update $\boldsymbol{z}$: Generate z^* using the Gibbs sampler from the conditional density

$f(z^*_{kl}|\boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}, \boldsymbol{y})$, where $\boldsymbol{z}_{-(kl)}$ denotes the set of all elements of $\boldsymbol{z}$ except for z_{kl}. In the Appendix, we show

$$Z^*_{kl}|\boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}, \boldsymbol{y} \sim N(E^2F, E^2), \tag{2.17}$$

where

$$\begin{aligned} E^2 =& \left\{ \frac{1}{\sigma^2} + \sum_{i=1}^{n} \frac{(\boldsymbol{r}_i^T Q(\boldsymbol{\beta}))^2_{kl}}{\tau^2 + \sigma_i^2} \right\}^{-1}, \\ F =& \sum_{i=1}^{n} \left[\frac{\{y(s_i) - \nu(s_i) - \alpha_{kl}(i)\}(\boldsymbol{r}_i^T Q(\boldsymbol{\beta}))_{kl}/\tau^2}{1 + \sigma_i^2/\tau^2} \right] \\ &+ \frac{1}{\sigma^2} \left\{ \beta_h \sum_{(k',l') \in \partial_{h(kl)}} z_{k'l'} + \beta_v \sum_{(k',l') \in \partial_{v(kl)}} z_{k'l'} + \beta_d \sum_{(k',l') \in \partial_{d(kl)}} z_{k'l'} \right\}, \end{aligned} \tag{2.18}$$

and $\left(\boldsymbol{r}_i^T Q(\boldsymbol{\beta})\right)_{kl}$ denotes the $((k-1) \times N + l)^{th}$ element of the vector $\boldsymbol{r}_i^T Q(\boldsymbol{\beta})$, $\alpha_{kl}(i) = \sum_{ab \neq kl} (\boldsymbol{r}_i^T Q(\boldsymbol{\beta}))_{ab} z_{ab}$, and $\partial_{h(kl)}, \partial_{v(kl)}, \partial_{d(kl)}$ denotes a horizontal, a vertical, and a diagonal neighborhood of (k, l) in W.

(2) Update θ_1: Generate θ_1^* using the MH algorithm from the conditional density

$$f(\theta_1|\theta_2, \theta_3, \boldsymbol{z}, \boldsymbol{y}) \propto f(\boldsymbol{z}|\theta_1)\pi(\theta_1)f(\boldsymbol{y}|\boldsymbol{z}, \theta_1, \theta_2, \theta_3).$$

(3) Update θ_2: Generate θ_2^* using the MH algorithm from the conditional density

$$f(\theta_2|\theta_1, \theta_3, \boldsymbol{z}, \boldsymbol{y}) \propto \pi(\theta_2)f(\boldsymbol{y}|\boldsymbol{z}, \theta_1, \theta_2, \theta_3).$$

(4) Update θ_3: Generate θ_3^* using the MH algorithm from the conditional density

$$f(\theta_3|\theta_1, \theta_2, \boldsymbol{z}, \boldsymbol{y}) \propto \pi(\theta_3|\theta_1)f(\boldsymbol{y}|\boldsymbol{z}, \theta_1, \theta_2, \theta_3).$$

It is easy to see that the time complexity (in each iteration) of the sampling algorithm for the model M_{AL} is $O(nMN)$. Since we usually set $MN \approx n$, the

complexity of the algorithm is $O(n^2)$. Since the number of iterations needed for a simulation to reach equilibrium does not significantly increase with the number of observations, the overall computational complexity of the model M_{AL} is still about $O(n^2)$. In Section D, we describe a fixed neighborhood version of the model M_{AL}, which can reduce the computational complexity to $O(n)$.

D. Bayesian Analysis for a Fixed Neighborhood Model

In the model M_{AL}, we have introduced an auxiliary lattice W to cover the region of observations, and define a GMRF Z on the auxiliary lattice to approximate the Gaussian process $\{X(s)\}$. The model M_{AL} allows $X(s_i)$ to depend on all components of Z. Since Z itself forms a GMRF, it is reasonable to assume that $X(s_i)$ is only dependent on a fixed subset of Z based on the observation of Rue and Tjelmeland (2002). This subset can be defined as follows.

Let W be an auxiliary lattice of size $M \times N$. For convenience, we denote by $s^w_{kl} = (s^w_k, s^w_l) \in \mathbb{R}^2$ the geographical site of the grid point (k, l) of W, and denote by $\partial_{kl} \subset W$ the neighborhood of s^w_{kl} in W. For any point $s \in \mathbb{R}^2$, we denote by $w(s)$ the nearest grid point to s; that is,

$$w(s) = \arg\min_{s^w_{kl} \in W} \|s - s^w_{kl}\| .$$

Define ∂s to be the neighboring set of $w(s)$ in the auxiliary lattice; that is, $\partial s = \partial w(s)$.

Consider the second-order neighboring structure of W. If $w(s) = (k, l)$, then ∂s can be written in the form of matrix by

$$\partial s = \partial_{kl} = \begin{bmatrix} (k-1, l-1) & (k-1, l) & (k-1, l+1) \\ (k, l-1) & (k, l) & (k, l+1) \\ (k+1, l-1) & (k+1, l) & (k+1, l+1) \end{bmatrix},$$

and the joint density $f\{x(s_1), \ldots, x(s_n)|\boldsymbol{z}\}$ can be written as

$$f\{x(s_1), \ldots, x(s_n)|\boldsymbol{z}\} = f\{x(s_1)|\ \overrightarrow{\boldsymbol{z}_{\partial s_1}}\} \cdots f\{x(s_n)|\ \overrightarrow{\boldsymbol{z}_{\partial s_n}}\}, \tag{2.19}$$

where $\overrightarrow{\boldsymbol{z}_{\partial s_i}} = \begin{bmatrix} z_{(\partial s_i)_{11}} & z_{(\partial s_i)_{12}} & \cdots & z_{(\partial s_i)_{33}} \end{bmatrix}^T$ denotes the values of Z_{ij} in the neighboring set of s_i.

We assume $\{X(s_i),\ \overrightarrow{Z_{\partial s_i}}\}$ is distributed as a multivariate Gaussian distribution with mean zero and the covariance matrix given by

$$\Sigma_{i,\partial} = \sigma^2 \begin{bmatrix} 1 & \boldsymbol{r}_{i,\partial}^T \\ \boldsymbol{r}_{i,\partial} & Q_\partial(\boldsymbol{\beta})^{-1} \end{bmatrix}, \tag{2.20}$$

where $\boldsymbol{r}_{i,\partial}$ is the correlation coefficient between $X(s_i)$ and $\overrightarrow{Z_{\partial s_i}}$, and $Q_\partial(\boldsymbol{\beta})^{-1}$ is the correlation matrix of $\overrightarrow{Z_{\partial s_i}}$. It follows from (2.2) that

$$Q_\partial(\boldsymbol{\beta}) = I_3 \otimes (I_3 - \beta_h S_3) + S_3 \otimes (-\beta_v I_3 - \beta_d S_3).$$

Therefore,

$$X(s_i)|Z(\partial s_i) \sim N\left\{\boldsymbol{r}_{i,\partial}^T Q_\partial(\boldsymbol{\beta})\ \overrightarrow{Z_{\partial s_i}}, \sigma^2\left(1 - \boldsymbol{r}_{i,\partial}^T Q_\partial(\boldsymbol{\beta})\, \boldsymbol{r}_{i,\partial}\right)\right\}. \tag{2.21}$$

For this fixed neighborhood system, the conditional density $f(z_{kl}^*|\boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}, \boldsymbol{y})$ is given by

$$Z_{kl}^*|\boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}, \boldsymbol{y} \sim N(E_\partial^2 F_\partial, E_\partial^2), \tag{2.22}$$

where

$$
\begin{aligned}
E_{\partial}^{2} &= \left\{ \frac{1}{\tau^2} + \sum_{i=1}^{n} I\left(z_{kl}^{*} \in \partial s_i\right) \frac{\left(\boldsymbol{r}_{i,\partial}^{T} Q_{\partial}(\boldsymbol{\beta})\right)_{i(kl)}^{2}}{\tau^2 + \sigma_i^2} \right\}^{-1}, \\
F_{\partial} &= \sum_{i=1}^{n} I\left(z_{kl}^{*} \in \partial s_i\ \right) \left[\frac{\left\{ y(s_i) - \nu(s_i) - \alpha_{kl}(i) \right\} \left(\boldsymbol{r}_{i,\partial}^{T} Q_{\partial}(\boldsymbol{\beta})\right)_{i(kl)} / \tau^2}{1 + \sigma_i^2/\tau^2} \right] \\
&\quad + \frac{1}{\tau^2} \left\{ \beta_h \sum_{(k',l') \in \partial_{h(kl)}} z_{k'l'} + \beta_v \sum_{(k',l') \in \partial_{v(kl)}} z_{k'l'} + \beta_d \sum_{(k',l') \in \partial_{d(kl)}} z_{k'l'} \right\},
\end{aligned}
\tag{2.23}
$$

where $i(kl)$ denotes that z_{kl}^{*} is the $i(kl)^{th}$ element of $\overrightarrow{Z_{\partial s_i}}$, $\left(\boldsymbol{r}_{i,\partial}^{T} Q_{\partial}(\boldsymbol{\beta})\right)_{i(kl)}$ denotes the $i(kl)^{th}$ element of the vector $\boldsymbol{r}_{i,\partial}^{T} Q_{\partial}(\boldsymbol{\beta})$, $\alpha_{kl}(i) = \sum_{ab \in \partial s_i, ab \neq kl} \left(\boldsymbol{r}_{i,\partial}^{T} Q_{\partial}(\boldsymbol{\beta})\right)_{i(ab)} z_{ab}$, and $\partial_{h(kl)}, \partial_{v(kl)}, \partial_{d(kl)}$ denotes a horizontal, a vertical, and a diagonal neighborhood in W of (k, l). For convenience, we will, henceforth, call the new model the *fixed neighborhood auxiliary lattice Gaussian* model and denote it by M_{FAL} in a shorthand notation.

Since the neighborhood size of M_{FAL} is fixed, it is easy to see that the time complexity (in each iteration) of the sampling algorithm (given in Section C) is $O(MN) + O(n)$, where the first term is for imputing the GMRF Z and the second term is for likelihood evaluation in drawing samples of $\boldsymbol{\theta}$. If $MN \approx n$, then the computational complexity is $O(n)$. Since the number of iterations needed for a simulation to reach equilibrium does not significantly increase with the number of observations, the overall computational complexity of the model M_{FAL} is still about $O(n)$. In this Chapter, for all datasets we tried with the sample size ranging from 1,000 to 12,000, the number of iterations is set to be about 30,000.

We note that even the neighborhood size is fixed, M_{FAL} can still approximate Gaussian random fields very well even for those with long correlation lengths. This will be illustrated in Section E. In addition, we can expect that M_{FAL} will cost much

less CPU times than M_G. This will also be illustrated in Section E.

E. Simulation Studies

In this section, we assess the performance of the models M_{AL} and M_{FAL} using simulated examples along with comparisons with the model M_G. Throughout all simulations of this Chapter, we set $\beta_h = \beta_v = \beta_d = \beta$, and set $L_{\sigma^2} = 0.01v_s^2$, $U_{\sigma^2} = 3v_s^2$, $L_{\tau^2} = 0.01v_s^2$, and $U_\phi = 10^6$, where v_s^2 denotes the variance of observations.

1. Model Estimation

We simulated 10 independent datasets from each of the models M_{AL} and M_{FAL} with the parameters $(\xi_0, \sigma^2, \tau^2) = (1, 3, 1)$ under each setting of $(\boldsymbol{\beta}, \phi)$ given in Table I. To simulate the data, we first simulate an auxiliary GMRF Z on a 30×30 lattice of grid size, and then draw $n = 1,000$ observation sites, $s_1, \ldots, s_{1000}$, uniformly on the region $[0, 100] \times [0, 100]$. Finally, we generate $\{X(s_1), \ldots, X(s_{1000})\}$ and $\{Y(s_1), \ldots, Y(s_{1000})\}$ according to equations (2.6) and (1.1) for the model M_{AL} and (2.21) and (1.1) for the model M_{FAL}. To re-estimate the parameters of the model, the Metropolis-within-Gibbs sampler, described in Section C, was used. For each dataset, the algorithm was run for 30,000 iterations and 1,000 samples were collected from the last 20,000 iterations at equally-spaced time points. The resulting parameter estimates are given in Table I.

The numerical results indicate the validity of our sampling scheme for both models, each parameter being correctly estimated.

Table I. Parameter estimation of the models M_{AL} and M_{FAL} for the simulated data. The numbers in the parentheses denote the standard deviations of the averaged estimates (over 10 datasets).

True	*Bias*				
Model (β, ϕ)	β	ϕ	ξ_0	σ^2	τ^2
M_{AL} (0.10, 8.0)	0.005(0.006)	-0.074(0.384)	-0.005(0.075)	0.209(0.374)	-0.237(0.201)
M_{AL} (0.10, 9.5)	-0.003(0.002)	0.025(0.422)	-0.018(0.061)	-0.172(0.492)	0.104(0.221)
M_{FAL} (0.12, 8.0)	0.008(0.004)	0.026(0.409)	0.011(0.062)	0.076(0.449)	-0.167(0.298)
M_{FAL} (0.12, 9.5)	-0.003(0.004)	-0.453(0.322)	-0.025(0.100)	0.245(0.481)	-0.178(0.193)

2. Approximation to Gaussian Random Fields

In this section, we assess the approximation ability of M_{AL} and M_{FAL} to Gaussian random fields. This is done under two scenarios, which are for short and long correlation lengths, respectively. The result for the long correlation range case is reported in Appendix. For the short range case, we simulated 10 independent datasets of size 1,500 from the model M_G, for which the spherical correlation function is used with the parameter $\phi = 20$. For each dataset, the locations $s_1, \ldots, s_{1500}$ were drawn uniformly from the region $[0, 100] \times [0, 100]$, and the observations $\{Y(s_1), \ldots, Y(s_{1500})\}$ were simulated with the parameters $(\xi_0, \phi, \sigma^2, \tau^2) = (1, 20, 7, 1)$ using the function grf() in geoR (Ribeiro and Diggle, 2001). A subset of 1,000 samples were randomly selected from the 1,500 samples and used for model estimation, and the rest 500 samples were used for prediction.

The model M_G was first applied to this dataset. The simulation was done using the function krige.bayes() in geoR (Ribeiro and Diggle, 2001). For this model, we

adopted a flat prior for ξ_0; a reciprocal prior for ϕ with the default discrete support set, 51 values equally spaced between 0 and 2 times the maximum distance between the data locations; and an uniform prior for τ^2/σ^2 with the default discrete support of 20 points in $(0, 1)$. In order to save CPU times, geoR has been implemented by restricting the sample spaces of ϕ and τ^2/σ^2 to a finite number of points. However, even with this restriction, as shown in Table II, it still consumes much longer CPU times than the model M_{FAL}.

Table II indicates that the parameters of the model M_G can be correctly estimated using *geoR*. Hence, the resulting mean squared prediction error (MSPE) can be used as a benchmark value for assessing the prediction ability of the models M_{AL} and M_{FAL}.

For the model M_{AL}, we considered a lattice of size 20×20. For each dataset, the Metropolis-within-Gibbs sampler was run for 23,000 iterations, for which the first 3,000 iterations were discarded for the burn-in process and 1,000 samples were collected from the remaining iterations at equally-spaced time points. The resulting parameter estimates and MSPEs are reported in Table II. For the model M_{FAL}, we consider four lattice sizes, 20×20, 30×30, 40×40 and 50×50. For each dataset and each lattice size, the Metropolis-within-Gibbs was run for 21,000 iterations, with the first 1,000 iterations being discarded for the burn-in process and then 1,000 samples being collected from the remaining iterations at equally-spaced time points. The resulting parameter estimates and MSPEs are also reported in Table II. Note that we usually set a little longer burn-in time for the model M_{AL} than the model M_{FAL}, as the GMRF is more dependent in M_{AL}.

To have a further exploration for the prediction performance, we have investigated the scatter plots of predicted values versus true values for the test samples of the 10 datasets. The plot shows no difference in predictions for the models M_G, M_{AL}

and M_{FAL}. However, as shown in Table II, M_{FAL} costs much shorter CPU times than the models M_{AL} and M_G.

Table II. Parameter estimation of the models M_{AL}, M_{FAL}, and M_G for the simulated data with a short correlation length of $\phi = 20$. The number in the parentheses denotes the standard deviation of the estimate. CPU: Measured in minutes on a 3.0 GHz Intel Core 2 Duo computer for a single run on one dataset.

	M_{AL}	M_{FAL}				M_G
	20×20	20×20	30×30	40×40	50×50	
β	.113(0.009)	0.111(0.004)	0.121(0.001)	0.123(0.000)	0.124(0.000)	
ϕ	15.86(1.97)	13.86(0.55)	9.73(0.29)	7.34(0.11)	6.01(0.03)	20.56(1.92)
ξ_0	1.09(0.37)	1.10(0.35)	1.11(0.36)	1.11(0.36)	1.10(0.37)	1.06(0.36)
σ^2	6.89(0.99)	6.32(0.54)	4.24(0.39)	3.20(0.34)	2.64(0.24)	7.17(1.02)
τ^2	0.51(0.16)	0.61(0.15)	0.90(0.19)	0.91(0.14)	0.92(0.08)	1.00 (0.09)
MSPE	2.37(0.17)	2.35(0.15)	2.26(0.13)	2.25(0.12)	2.26(0.10)	2.18(0.11)
CPU(m)	99.41	2.28	2.76	2.54	2.77	101.7

Regarding parameter estimation of the three models, we have a few remarks:

- **Remark 1.** Due to the use of auxiliary lattices, the parameters of M_{AL} and M_{FAL} are not directly comparable with the parameters of M_G, except for the mean parameter ξ_0. Table II shows that ξ_0 can be correctly estimated in all the three models.

  For the models M_{AL} and M_{FAL}, although the estimates of ϕ and σ^2 are different for different lattice sizes, their ratios are relatively stable. Based on the estimates given in Table II, we have $\hat{\phi}/\hat{\sigma}^2$=2.2, 2.2, 2.3, 2.3, and 2.3 for the models

M_{AL} and M_{FAL}'s (from left to right). The size of the auxiliary lattice can have a significant effect on parameter estimation, particularly on β, ϕ and σ^2. This is reasonable: When a larger lattice is used, the dependence between neighboring grid points becomes stronger, the correlation length and the marginal variance σ^2 can then be reduced accordingly. However, the prediction performance is not significantly affected by the lattice size.

- **Remark 2.** As the size of auxiliary lattice approaches to the number of observations, M_{FAL} can achieve its best prediction performance, closing to that of the true model M_G from which the data were generated. When the lattice size is much larger than the number of observations, the prediction performance of M_{FAL} may deteriorate due to its fixed neighborhood structure. In practice, the size of auxiliary lattice can be determined using a cross-validation approach. Alternative to the cross-validation approach, a more Bayesian method, the DIC method (Spiegelhalter *et al.*, 2002), can also be applied to determine the size of auxiliary lattice.

- **Remark 3.** When the auxiliary lattices are the same, the models M_{AL} and M_{FAL} can perform very similarly in prediction (see Table II). The main difference between the two models is at CPU time. Since the model M_{AL} allows full dependence of observations in model fitting, it can cost much longer CPU time than the model M_{FAL}. Recall that the computational complexity of M_{AL} is $O(n^2)$, while the computational complexity of M_{FAL} is only $O(n)$, provided that $MN \approx n$ holds.

- **Remark 4.** For a fixed dataset, the CPU time cost by M_{FAL} does not necessarily increase with the size of auxiliary lattice, although it tends to do so. The reason is that in simulating the GMRF Z, the number of observations fallen

into the neighborhood of each component of Z tends to decrease as the size of auxiliary size increases, and this can reduce the computational time for E_∂^2 and F_∂ given in (2.23). This explains why M_{FAL} costs shorter CPU time with a a 40×40-lattice than with a 30×30-lattice (shown in Table II).

3. Large Sample Study

To investigate the scalability of M_{FAL}, we conducted two experiments under the settings given in Section 2 with $\phi = 20$ and $\phi = 40$. Under each setting, we simulated 10 datasets of size $n = 4,500$ from the model M_G. The locations of the samples were uniformly distributed on $[0, 100] \times [0, 100]$. For each data set, 3,000 randomly selected samples were used for model estimation, and the rest were used for prediction. The model M_{FAL} was applied to this example. The lattice sizes we considered are 30×30, 40×40 and 50×50. For each data set, the algorithm was run for 21,000 iterations, for which the first 1,000 iterations were used for the burn-in process and then every 20^{th} sample was collected from the remaining iterations. The numerical results were summarized in Table III. For these datasets, the model M_G was not applicable as the training set is too large.

The numerical results indicate that for large data sets, even when the model M_G does not work, the model M_{FAL} can still work very well. When the lattice size is close to the number of training samples, M_{FAL} can produce very good prediction results.

4. CPU and Memory Complexity Analysis

To evaluate computational complexity of the model M_{FAL}, we measured the CPU time and memory space cost by it for different sample sizes when the lattice size exactly matches the sample size. For the memory space, we measured the resident set size (RSS), which is the amount of physical memory mapped into the process

Table III. Summary of estimation, prediction, and CPU times of the model M_{FAL} for large data sets. The number in the parentheses denotes the standard deviation of the estimate. CPU: Measured in minutes on a 3.0 GHz Intel Core 2 Duo computer for a single run for one dataset.

	$\phi = 20$			$\phi = 40$		
	30×30	40×40	50×50	30×30	40×40	50×50
β	0.119(0.001)	0.124(4e-4)	0.124(1e-4)	0.124(3e-4)	0.124(9e-5)	0.124(4e-5)
ϕ	9.55(0.25)	7.41(0.06)	5.92(0.04)	10.02(0.20)	7.50(0.05)	5.93(0.04)
ξ_0	0.98(0.38)	0.98(0.39)	0.98(0.39)	1.01(0.54)	1.00(0.54)	0.99(0.54)
σ^2	4.26(0.31)	2.94(0.21)	2.29(0.16)	2.28(0.26)	1.60(0.18)	1.35(0.13)
τ^2	0.89(0.12)	1.08(0.05)	1.10(0.05)	0.95(0.06)	1.01(0.05)	0.98(0.04)
MSPE	1.88(0.04)	1.84(0.05)	1.82(0.05)	1.50(0.06)	1.47(0.05)	1.47(0.05)
CPU(m)	6.59	6.59	6.90	6.61	6.42	6.55

during simulations.

We consider four sample sizes 400, 900, 1600, and 2,500 with the respective lattice size 20×20, 30×30, 40×40, and 50×50. For each case, MCMC was run for 5,000 iterations, and the CPU time and RSS were measured. The results are summarized in Table IV. For CPU times, we fit an exponential function $\text{CPU}(n) = 0.0413n^{0.9679}$; and for memory space, we fit a linear function $\text{Memory}(n) = 0.0008831n + 0.9777$. The fitted functions are shown in Figure 2. This indicates that both the CPU time and memory space cost by the model M_{FAL} is a linear function of the sample size. Therefore, M_{FAL} can be applied to very large dataset.

Table IV. Computational complexity of the model M_{FAL}. CPU: measured in seconds on a 3.0 GHz Intel Core 2 Duo computer for a single run of M_{FAL}; Memory: resident set size measured in MB during simulations.

Sample size	CPU(s)	Memory(MB)
400	13.47	1.31
900	30.46	1.84
1600	52.51	2.32
2500	79.28	3.21

F. Geostatistical Data Study

In this section, we study the performance of the model M_{FAL} on two real datasets, along with a comparison with the model M_G. Our numerical results show that for real datasets, our model can generally outperforms the model M_G in both prediction errors and CPU times. In addition, our model can be applied to very large datasets.

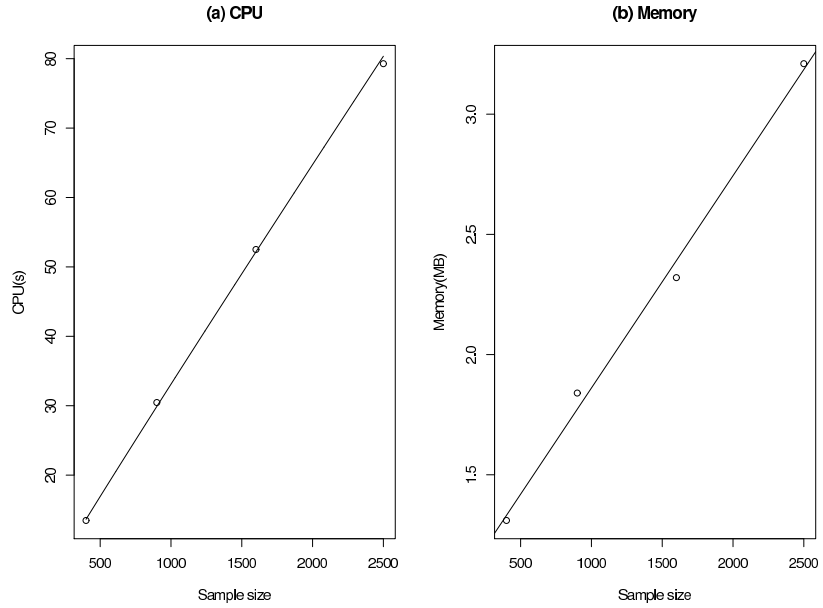

Fig. 2. Computational complexity of the model M_{FAL}. (a) CPU time: the fitted function is $\text{CPU}(n) = 0.0413n^{0.9679}$; (b) Memory: the fitted function is $\text{Memory}(n) = 0.0008831n + 0.9777$.

1. Geevor Data

The Geevor data is a sample data from a hydrothermal tin deposit in Cornwall, England. Ore was extracted from the underground lodes by overhead shrinkage stopping, between horizontal development drives (on lode) approximately 100 feet apart (Clark and Harper, 2000). The lode is sampled by chipping across the vein in the hanging wall of the drive or the stope. Samples of around 1kg are chipped across the vein, which averages about 24 inches wide. Measurements are grades of tin in pounds of black tin (SnO2) per ton of rock. The thickness of the vein or lode is measured to the nearest inch. Coordinates are in feet along section and elevation above an arbitrary base level.

The data consists of samples in stope and development. To compare with the model M_G, we consider only a subset of the data, the stope samples located above 600 of y-coordinates (shown in Figure 3). Since there are two locations at which we had two observations, we excluded them and the resulting dataset consisted of 1,242 observations. We randomly selected 1,000 observations for model estimation and used the rest 242 observations for prediction.

The model M_{FAL} was applied to this example with an auxiliary lattice of 20×80, which covers the region $[1357.0, 3005.0] \times [602.0, 904.0]$. The MCMC was run 10 times with different initial values. Each run consisted of 51,000 iterations with the first 1,000 iterations being discarded for the burn-in process. The samples were then collected at every 50^{th} iterations from the remaining iterations. The numerical results were summarized in Table V.

Since the sample size 1,000 is still manageable for the model M_G, we applied it to this example with three different correlation functions, spherical, exponential, and Matérn with $\kappa = 1$. For this model, we adopted a flat prior for ξ_0, a reciprocal

prior for ϕ with the default discrete support, and a uniform prior for τ^2/σ^2 with a discrete support of 21 points equally spaced in $(0, 0.5)$. The numerical results were summarized in Table V. The comparison shows that for this example the model M_{FAL} significantly outperforms the model M_G in terms of predictions. In addition, it costs much less CPU times than the model M_G.

Since the MSPE of M_G is much worse than that of the M_{FAL}, to evaluate the validity of the M_G for this dataset, we checked the isotropy assumption through an exploratory data analysis. The semivariograms of the training samples in the directions of 0°, 45° ,90° ,135° are depicted in Figure 3 using the command *variog4()* in geoR. The plot shows that the semivariograms slightly depend on the chosen directions, and so the data slightly violates the assumption of isotropy. This example indicates that the auxiliary lattice model is more robust than the M_G model to possible violation of isotropy of the data. Possible extensions of the auxiliary lattice model for handling anisotropy or non-stationarity are discussed in Chapter IV.

2. Goldmine Samples

This dataset is constructed based on a Wits type gold mine some decades into production. The samples are chipped from the face of the reef in a working section of the mine (stope). As the face advances, new chip samples are taken. Values within a stope are traditionally estimated using the sample values from the face. The dataset is available at `http://www.kriging.com/datasets/`.

To have the data region closer to a rectangular, we first rotated the locations of the samples 33.2° clockwisely, and then selected all samples falling into the rectangular region $[300, 3300] \times [500, 1200]$. This resulted in 14,932 samples being selected. We then randomly selected 12,000 samples for model estimation and used the remaining 2,932 samples for model assessment. The model M_{FAL} was applied to this example

Table V. Parameter estimation of M_{FAL} for the Geevor data. The number in the parentheses denotes the standard deviation of the estimates. CPU: Measured in minutes on a 3.0 GHz Intel Core 2 Duo computer for a single run of the corresponding model.

	M_{FAL}	M_G		
	20×80	Spherical	Exponential	Matérn $\kappa = 1$
β	0.091(0.004)			
ϕ	37.01(0.46)	1815.88	1288.86	317.95
ξ_0	43.99(0.10)	19.28	22.36	16.24
σ^2	2090.49(13.05)	10027.61	10681.55	10145.47
τ^2	1832.89(11.10)	3538.16	3688.74	3966.33
MSPE	3395.05(8.25)	3552.49	3550.77	3613.30
CPU(m)	5.7	283.3	266.7	148.3

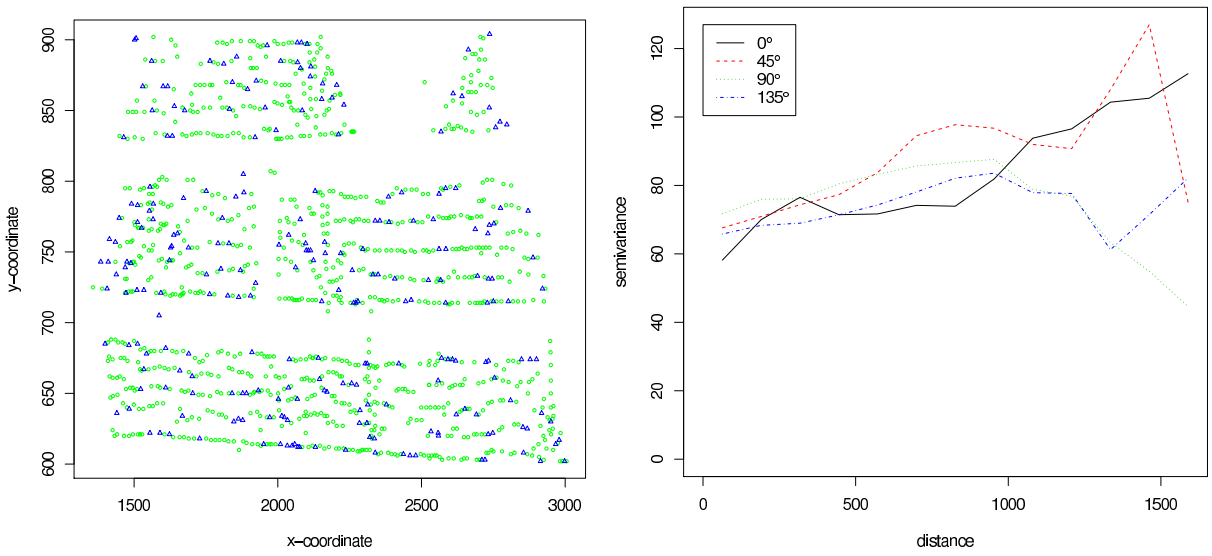

Fig. 3. Left: Plot of the 1,000 sites used for estimation (circle) and the 242 sites (triangle) used for prediction. Right: Empirical variograms of the training data with angle 0°,45°,90°,135°.

with a lattice size of 50×200 covering the region $[300, 3300] \times [500, 1200]$. To fit Gaussian models to this data, the observations are logarithmically transformed. A QQ-plot (omitted in the dissertation) shows that the transformed data are approximately normally distributed. The algorithm was run 5 times with different initial values. Each run consisted of 32,000 iterations. We discarded the first 2,000 iterations for the burn-in process and then collected the samples at every 30^{th} time points. The numerical results are summarized in Table VI. Figure 5 shows the images of the observations and predicted values at the testing sites. The comparison implies that our method is of practical value: It is very good in prediction, although it is very fast.

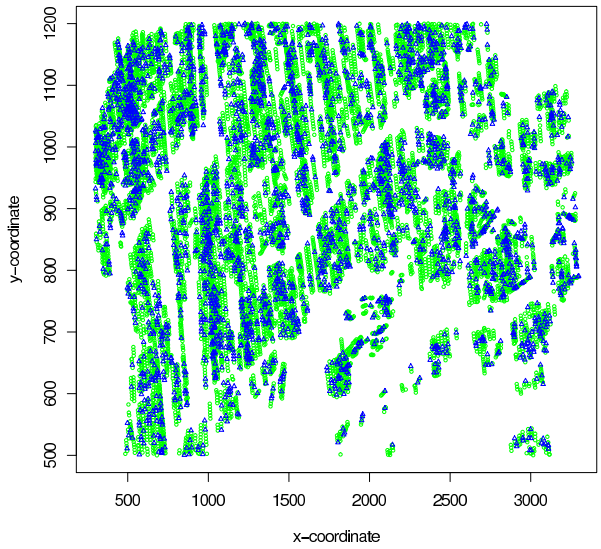

Fig. 4. Scatter plot of the 12,000 locations (denoted by circle) used for model estimation and the 2,932 locations (denoted by triangle) used for prediction.

Table VI. Estimation results of the model M_{FAL} for the large sample data. The number in the parentheses denotes the standard deviation of the estimate. CPU: Measured in minutes on a 3.0 GHz Intel Core 2 Duo computer for a single run of M_{FAL}.

	50×200
β	0.1162(0.0002)
ϕ	33.683(0.0346)
ξ_0	3.742 (0.0004)
σ^2	0.365 (0.0004)
τ^2	0.011 (0.0005)
MSPE	0.152 (0.0001)
CPU(m)	42

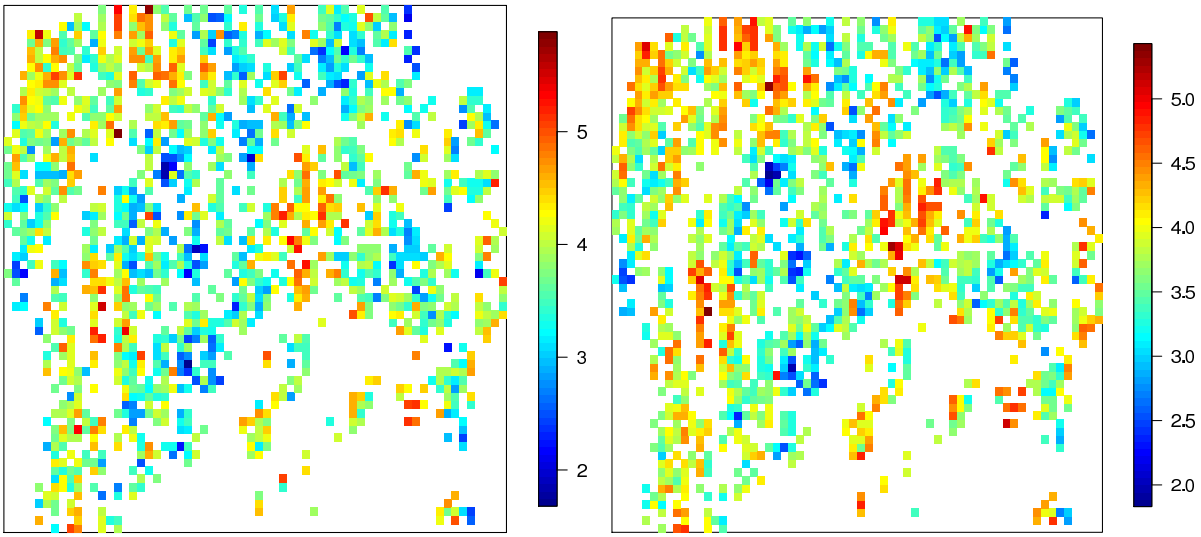

Fig. 5. Left : Images of observations at the testing sites. Right : Images of predicted values at the testing sites.

CHAPTER III

A PREDICTION-ORIENTED BAYESIAN SITE SELECTION APPROACH

A. The Regression Model Formulation

Let $D = \{y(s_i)\}$ denote the observations drawn from the model (1.1) at at n distinct locations $\boldsymbol{s} = \{s_1, \cdots, s_n\}$, and let $\boldsymbol{s}^p = \{s_1^p, \cdots, s_{n_p}^p\}$ denote n_p distinct locations of interest for prediction. Suppose that D has been partitioned two sets, $D_y = \{y(s_i); s_i \in \boldsymbol{s}^y, i = 1, \cdots, n^*\}$ and $D_{-y} = D \backslash D_y$, where $\boldsymbol{s}^y = \{s_1^y, \cdots, s_{n^*}^y\}$ is the set of locations of the observations contained in D_y. In addition, we assume that D_y has been selected to consist of all observations that are near the prediction sites $\boldsymbol{s}^p$. How to select D_y will be discussed in Section B.

Let $Y(\boldsymbol{s}^y) = \{Y(s_1^y), \cdots, Y(s_{n^*}^y)\}^T$ denote the vector of observations contained in D_y. Likewise, let $Z(\boldsymbol{s}^{-y})$ denote the vector of observations contained in D_{-y}. Following from model (1.1), the distribution of $Y(\boldsymbol{s}^y)$ conditioned on $Z(\boldsymbol{s}^{-y})$ follows a multivariate normal distribution; that is, a normal regression can then be formulated as

$$Y(\boldsymbol{s}^y) \sim Z(\boldsymbol{s}^{-y}),$$

where $Y(\boldsymbol{s}^y)$ works as the response variable and $Z(\boldsymbol{s}^{-y})$ works as the explanatory variable. Instead of using all $Z(\boldsymbol{s}^{-y})$ as explanatory variables, we would select a subset of $Z(\boldsymbol{s}^{-y})$ as the explanatory variables for $Y(\boldsymbol{s}^y)$, as the variables in $Z(\boldsymbol{s}^{-y})$ can be highly correlated given the nature of spatial model (1.1). With a little abuse of notations, we denote by $Z = \{Z(s_1^z), \ldots Z(s_m^z)\}$ the set of variables used as the explanatory variables of $Y(\boldsymbol{s}^y)$, where $m = |Z|$ denotes the size of the set Z. Then the conditional distribution $[Y(\boldsymbol{s}^y)|Z]$ is given by

$$Y(\boldsymbol{s}^y)|Z \sim N\left(\boldsymbol{\nu}_{y|z}, \Sigma_{y|z}\right) \tag{3.1}$$

where

$$\begin{aligned} \boldsymbol{\nu}_{y|z} &= \boldsymbol{\nu}_y + \Sigma_{yz}\Sigma_z^{-1}\left(Z - \boldsymbol{\nu}_z\right), \\ \Sigma_{y|z} &= \Sigma_y - \Sigma_{yz}\Sigma_z^{-1}\Sigma_{zy}. \end{aligned} \tag{3.2}$$

Let $R_y = Corr\{Y(\boldsymbol{s}^y)\}$ denote the correlation matrix of $Y(\boldsymbol{s}^y)$, let $R_z = Corr(Z)$ denote the correlation matrix of Z, and let $R_{yz} = Corr\{Y(\boldsymbol{s}^y), Z\}$ denote the correlation matrix between $Y(\boldsymbol{s}^y)$ and Z. Then the covariance matrices in (3.2) can be expressed as

$$\Sigma_y = \sigma^2\{R_y(\phi) + \alpha I\}, \quad \Sigma_z = \sigma^2\{R_z(\phi) + \alpha I\}, \quad \Sigma_{yz} = \sigma^2 R_{yz}, \quad \Sigma_{yz} = \Sigma_{zy}^T,$$

where $\alpha = \tau^2/\sigma^2$.

In the case of covariates presented in model (1.1), we have

$$\boldsymbol{\nu}_{y|z} = \boldsymbol{\nu}_y + \Sigma_{yz}\Sigma_z^{-1}\left(Z - \boldsymbol{\nu}_z\right) = \left(C_y - R_{yz}R_z^{-1}C_z\right)\boldsymbol{\xi} + R_{yz}R_z^{-1}Z, \tag{3.3}$$

where $\boldsymbol{\xi} = (\xi_0, \xi_1, \cdots, \xi_p)^T$ denotes the vector of regression coefficients as defined in (1.3), and C_y and C_z are the design matrices for the covariate and given by

$$C_y = \begin{bmatrix} 1 & c_{s_1^y,1} & \cdots & c_{s_1^y,p} \\ \vdots & & \vdots & \\ 1 & c_{s_{n^*}^y,1} & \cdots & c_{s_{n^*}^y,p} \end{bmatrix},$$

$$C_z = \begin{bmatrix} 1 & c_{s_1^z,1} & \cdots & c_{s_1^z,p} \\ \vdots & & \vdots & \\ 1 & c_{s_m^z,1} & \cdots & c_{s_m^z,p} \end{bmatrix}.$$

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \sigma^2, \boldsymbol{\xi}) = \{\log(\phi), \log(\alpha), \sigma^2, \boldsymbol{\xi}\}$ denote the parameters of the model (3.1)-(3.3), where ϕ and α has been reparameterized by their logarithms. To make Bayesian inference for the model (3.1)-(3.3), we specify the following priors for

$\boldsymbol{\xi}$, σ^2 and ϕ:

$$\begin{aligned} \pi(\boldsymbol{\xi}|\sigma^2) &\propto \epsilon_\xi^{1+p}\sigma^{-(1+p)} \exp\left(-\frac{\epsilon_\xi^2}{2\sigma^2}\boldsymbol{\xi}^T\boldsymbol{\xi}\right), \\ \pi(\sigma^2) &\propto IG(\epsilon,\epsilon), \\ \pi(\phi) &\propto IG(\epsilon,\epsilon), \end{aligned} \tag{3.4}$$

where both ϵ_ξ and ϵ are small positive constants, and $\mathrm{IG}(\cdot,\cdot)$ denotes an inverse Gamma distribution. For simplicity, the two hyperparameters of the prior inverse Gamma distribution are restricted to be the same in this Chapter. When $\epsilon \leq 2$, $\mathrm{IG}(\epsilon,\epsilon)$ leads to a vague prior, whose variance is infinite.

Since it is generally true that the nugget variance τ^2 is smaller than the variance σ^2, we set a uniform prior for $\alpha = \tau^2/\sigma^2$ on the interval $[0,1]$; that is,

$$\pi(\alpha) = 1, \quad \alpha \in [0,1]. \tag{3.5}$$

With a little abuse of notations, we denote the model (3.1) by Z and impose a truncated Poisson distribution on the space of models; that is,

$$\pi(Z) \propto \frac{\lambda^m}{m!}e^{-\lambda}, \quad m \in \{0,1,\ldots,n-n^*\}, \tag{3.6}$$

where $m = |Z|$ denotes the number of sites included in Z and λ is a hyperparameter to be pre-specified by the user.

Combining (3.2)–(3.3) and (3.4)–(3.6), we have the posterior of $\boldsymbol{\theta}$ given by

$$\begin{aligned} f\{\boldsymbol{\theta}|Y(\boldsymbol{s}^y),Z\} \propto \left|\Sigma_{y|z}\right|^{-1/2} \frac{1}{\sigma^{1+p}} \exp\left\{-\frac{1}{2\sigma^2}B^T(R_{y|z}+\epsilon_\xi^{-2}AA^T)^{-1}B\right\}\pi(\theta_1,\theta_2,\sigma^2) \\ \times \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\xi}-\Lambda^{-1}E)^T\Lambda(\boldsymbol{\xi}-\Lambda^{-1}E)\right\}, \end{aligned} \tag{3.7}$$

where $A = C_y - R_{yz}R_z^{-1}C_z$, $B = Y(\boldsymbol{s}^y) - R_{yz}R_z^{-1}Z$, $R_{y|z} = \sigma^{-2}(\Sigma_y - \Sigma_{yz}\Sigma_z^{-1}\Sigma_{zy})$

denotes the conditional correlation matrix of $Y(\boldsymbol{s}^y)$ given Z, $E = A^T R_{y|z}^{-1} B$ and $\Lambda = A^T R_{y|z}^{-1} A + \epsilon_\xi^2 I$ is an $n^* \times n^*$ matrix. It is worth pointing out that both $\Sigma_{y|z}$ and $R_{y|z} + \epsilon_\xi^{-2} AA^T$ are also $n^* \times n^*$ matrix. Thus, BSS reduces the problem of inverting $n \times n$ matrices to that of inverting $n^* \times n^*$ matrices. How to determine the value of n^* will also be discussed in Section B.

Integrating out $\boldsymbol{\xi}$ and σ^2 from (3.7), we have

$$f(\theta_1, \theta_2 | Y(\boldsymbol{s}^y), Z) \propto \left|R_{y|z}\right|^{-1/2} |\Lambda|^{-1/2} \frac{\Gamma\left(\frac{n}{2} + \epsilon\right)}{\left\{B^T (R_{y|z} + \epsilon_\xi^{-2} AA^T)^{-1} B/2 + \epsilon\right\}^{\frac{n}{2}+\epsilon}} \pi(\theta_1, \theta_2). \tag{3.8}$$

Following the standard theory of Bayesian model averaging, the predictive posterior distribution of $Y(\boldsymbol{s}^p)$ can be written as

$$f\{Y(\boldsymbol{s}^p) | Y(\boldsymbol{s}^y), D_{-y}\} = \sum_{Z \subset D_{-y}} \int f\{Y(\boldsymbol{s}^p) | Y(\boldsymbol{s}^y), Z, \boldsymbol{\theta}\} f\{\boldsymbol{\theta} | Y(\boldsymbol{s}^y), Z\} \pi(Z) d\boldsymbol{\theta}, \tag{3.9}$$

where Z denotes any subset of D_{-y} and also a particular model defined in (3.1)–(3.3). This implies that the expectation of $Y(\boldsymbol{s}^p)$ conditioned on the full observations D is given by

$$E\{Y(\boldsymbol{s}^p) | Y(\boldsymbol{s}^y), D_{-y}\} = \sum_{Z \subset D_{-y}} \int E\{Y(\boldsymbol{s}^p) | Y(\boldsymbol{s}^y), Z, \boldsymbol{\theta}\} f\{\boldsymbol{\theta} | Y(\boldsymbol{s}^y), Z\} \pi(Z) d\boldsymbol{\theta}. \tag{3.10}$$

Let $(\boldsymbol{\theta}^{(1)}, Z^{(1)}), \ldots, (\boldsymbol{\theta}^{(N)}, Z^{(N)})$ denote a sequence of samples drawn from the joint posterior of $(\boldsymbol{\theta}, Z)$, which is proportional to $f\{\boldsymbol{\theta} | Y(\boldsymbol{s}^y), Z\} \pi(Z)$. Then $E\{Y(\boldsymbol{s}^p) | Y(\boldsymbol{s}^y), D_{-y}\}$ can be estimated by

$$\widehat{Y}(s^p) = \frac{1}{N} \sum_{i=1}^{N} E\{Y(s_p) | Y(\boldsymbol{s}^y), Z^{(i)}, \boldsymbol{\theta}^{(i)}\}, \tag{3.11}$$

where $E\{Y(s_p) | Y(\boldsymbol{s}^y), Z^{(i)}, \boldsymbol{\theta}^{(i)}\}$ is the conditional mean of $Y(s_p)$ given $Y(\boldsymbol{s}^y)$, the selected set of explanatory variables $Z^{(i)}$, and the parameter values $\boldsymbol{\theta}^{(i)}$. How to draw

samples from the joint posterior of $(\boldsymbol{\theta}, Z)$ will be discussed in Section C.

B. Prediction-Oriented Response Variable Selection

In this section, we consider a prediction-oriented selection scheme for $Y(\boldsymbol{s}^y)$ with an expectation that $\{Y(\boldsymbol{s}^y)\}$ plays surrogates for $\{Y(\boldsymbol{s}^p)\}$. The scheme consists of the following steps:

1. Let $\boldsymbol{s} = \{s_1, \ldots, s_n\}$ denote the full set of observation sites, and let $\boldsymbol{s}^p = \{s_1^p, \ldots, s_{n_p}^p\}$ denote the set of prediction sites.

   For $i = 1, \ldots n^p$, do the following sub-steps to identify the first tier of the nearest points to $\boldsymbol{s}^p$:

   (a) Draw a site s_i^p from the set $\boldsymbol{s}^p$ at random and without replacement.

   (b) Identify the nearest neighbor of s_i^p by setting

   $$s_{1,i}^y = \arg \min_{s \in \boldsymbol{s} \backslash \{s_{1,1}^y, \ldots, s_{1,i-1}^y\}} \|s - s_i^p\|.$$

   Set $\boldsymbol{s}_1^y = \{s_{1,1}^y, \ldots, s_{1,n^p}^y\}$.

2. Set $\boldsymbol{s} \leftarrow \boldsymbol{s} \backslash \boldsymbol{s}_1^y$ and repeat the substeps in step 1 to identify the second tier of the nearest points to $\boldsymbol{s}^p$. Denote the second tier neighboring set by $\boldsymbol{s}_2^y$.

   .....

k. Set $\boldsymbol{s} \leftarrow \boldsymbol{s} \backslash \boldsymbol{s}_{k-1}^y$ and repeat the substeps in step 1 to identify the k-th tier of the nearest points to $\boldsymbol{s}^p$. Denote the k-th tier neighboring set by $\boldsymbol{s}_k^y$.

The procedure outputs $\boldsymbol{s}^y = \cup_{j=1}^k \boldsymbol{s}_j^y$ as the set of response variables and $D_{-y} = \{s_1, \ldots, s_n\} \backslash \boldsymbol{s}^y$ as the set of explanatory variables.

In practice, the value of k, which determines the size of $\boldsymbol{s}^y$ ($n^* = kn^p$), can be determined through an examination of the fitting to $\{Y(\boldsymbol{s}^y)\}$ or its subset. For example, we can choose the value of n^* such that the mean squared fitting errors (MSFE) for the first tier neighboring sites are minimized among a few values of n^* under consideration. Our numerical results indicate that MSFE can provide a good guideline for selection of n^*. In our experience, when $k \geq 3$, BSS works very well irrespective of the size of the original dataset.

As shown in (3.7), BSS has reduced the problem of inverting $n \times n$ matrices to that of inverting $n^* \times n^*$ matrices. When n^p, the number of prediction points, is large, we suggest to divide $\boldsymbol{s}^p$ into several small subsets and then run BSS for each of them separately. For example, the subsets can be constructed by drawing from $\boldsymbol{s}^p$ through a sampling-without-replacement procedure. This helps us to keep n^* in a reasonable range, alleviating the heavy burden of computation caused by the cubic law of matrix inversion.

C. A Metropolis-within-Gibbs Sampling Scheme

In this section, we consider a Metropolis-within-Gibbs sampler (Müller, 1991) for drawing samples from the posterior

$$f\{\theta_1, \theta_2, Z|Y(\boldsymbol{s}^y)\} \propto f\{\theta_1, \theta_2|Y(\boldsymbol{s}^y), Z\}\pi(Z),$$

where Z indexes a subset model and $f\{\theta_1, \theta_2|Y(\boldsymbol{s}^y), Z\}$ is given in (3.8).

Let $(\theta_1^{(t)}, \theta_2^{(t)}, Z^{(t)})$ denote the sample generated at iteration t of the Markov chain. Let $m = |Z^{(t)}|$ denote the number of sites included in $Z^{(t)}$. To update $Z^{(t)}$, we consider three possible moves, "birth", "death" and "exchange" with the respective proposal

probabilities denoted by $q_{m,m+1}$, $q_{m,m-1}$ and $q_{m,m}$. In this Chapter, we set

$$q_{m_{\min},m_{\min}} = \frac{1}{3}, \quad q_{m_{\min},m_{\min}+1} = \frac{2}{3},$$
$$q_{m_{\max},m_{\max}} = \frac{1}{3}, \quad q_{m_{\max},m_{\max}-1} = \frac{2}{3},$$

$$q_{i,i+1} = q_{i-1,i} = q_{i,i} = \frac{1}{3}, \quad \text{for } \; m_{\min}+1 \leq i \leq m_{\max}-1 \; ,$$

where $m_{\min} = 0$ and $m_{\max} = n - n^*$. One iteration of the Metropolis-within-Gibbs sampler consists of the following steps:

- Draw $\theta_1^{(t+1)}$ from the conditional distribution $f\{\theta_1|\theta_2^{(t)}, Y(\boldsymbol{s}^y), Z\}$ using the Metropolis algorithm with a random walk Gaussian proposal. The variance of this propoal is denoted by $\sigma^2_{\theta_1}$ and will be given in the context of numerical studies.

- Draw $\theta_2^{(t+1)}$ from the conditional distribution $f\{\theta_2|\theta_1^{(t+1)}, Y(\boldsymbol{s}^y), Z\}$ using the Metropolis algorithm with a random walk Gaussian proposal. The variance of this propoal is denoted by $\sigma^2_{\theta_2}$ and will be given in the context of numerical studies.

- Draw $Z^{(t+1)}$.

  - (*Birth*) Randomly select z^* out of $D_{-y} \backslash Z^{(t)}$ and set $Z^* = Z^{(t)} \cup z^*$. Set $Z^{(t+1)} = Z^*$ with probability

$$\min\left\{1, \frac{f\{\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\boldsymbol{s}^y), Z^*\}\pi(Z^*)}{f\{\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\boldsymbol{s}^y), Z^{(t)}\}\pi(Z^{(t)})} \frac{n-n^*-m}{m+1} \frac{q_{m+1,m}}{q_{m,m+1}}\right\}.$$

    Otherwise, set $Z^{(t+1)} = Z^{(t)}$.

  - (*Death*) Randomly select z^* out of $Z^{(t)}$ and set $Z^* = Z^{(t)} \backslash z^\star$. Accept $\boldsymbol{z}^\star_{m-1}$ with probability

$$\min\left\{1, \frac{f\{\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\boldsymbol{s}^y), Z^*\}\pi(Z^*)}{f(\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\boldsymbol{s}^y), Z^{(t)})\pi(Z^{(t)})} \frac{m}{n-n^*-m+1} \frac{q_{m-1,m}}{q_{m,m-1}}\right\}.$$

Otherwise, set $Z^{(t+1)} = Z^{(t)}$.

– (*Exchange*) Randomly select $z^\star$ out of $D_{-y} \backslash Z^{(t)}$ and $z_u^\star$ out of $Z^{(t)}$. Set $Z^* = Z^{(t)} \cup \{z^*\} \backslash \{z_u^\star\}$ by exchanging $z^\star$ and $z_u^\star$. Accept $\boldsymbol{z}_m^\star$ with probability

$$\min \left\{ 1, \frac{f\{\theta_1^{(t+1)}, \theta_2^{(t+1)} | Y(\boldsymbol{s}^y), Z^*\}}{f\{\theta_1^{(t+1)}, \theta_2^{(t+1)} | Y(\boldsymbol{s}^y), Z^{(t)}\}} \right\}$$

Otherwise, set $Z^{(t+1)} = Z^{(t)}$.

Given a MCMC sample $(\theta_1^{(t)}, \theta_2^{(t)}, Z^{(t)})$, $\boldsymbol{\xi}^{(t)}$ and $\sigma^{2(t)}$ can drawn from the following distributions:

$$\boldsymbol{\xi}^{(t)} \sim N\left(\Lambda^{-1}E, \Lambda^{-1}\right), \quad \sigma^{2(t)} \sim IG\left\{n/2 + \epsilon, B^T(R_{y|z} + \epsilon_\xi^{-2} AA^T)^{-1}B/2 + \epsilon\right\},$$

which can be simply derived from (3.7) with Λ, W, A, B and $R_{y|z}$ as defined before. Given the samples $(\theta_1^{(t)}, \theta_2^{(t)}, Z^{(t)})$ and $(\boldsymbol{\xi}^{(t)}, \sigma^{2(t)})$, the prediction of $\{Y(\boldsymbol{s}^p)\}$ can then be simply done as in (3.11).

D. Simulation Studies

In this section, we assess the performance of BSS using two simulated examples along with some comparisons with the standard Bayesian method. For the simulated examples, we have the following common settings. In both data generation and posterior simulations, the correlation function is from the exponential family

$$\text{Corr}\{Y(s_i), Y(s_j)\} = \exp\left\{-\frac{\|s_i - s_j\|}{\phi}\right\},$$

where $\|.\|$ denotes the Euclidean norm. In posterior simulations, we set the hyperparameters $\epsilon_\xi = 0.01$ and $\epsilon = 1$. As previously explained, this leads to vague priors for $\boldsymbol{\xi}$, σ^2 and ϕ. For each dataset, BSS was run once with 10,000 iterations, with the first 5,000 iterations being discarded for the burn-in process and the remaining iterations

are thinned by 5 to get 1,000 samples.

1. An Illustrative Example

We simulated 30 independent data sets from the Gaussian geostatistical model (1.1). Each data set contains 1,100 observations with the sites uniformly distributed over the region $[0, 100] \times [0, 100]$. The data sets were generated using the function grf() in geoR (Ribeiro and Diggle, 2001) with the parameters $(\xi_0, \xi_1, \phi, \sigma^2, \tau^2) = (0.5, 1, 25, 1, 0.25)$ and the covariates generated from $N(0, 1)$. For each data set, a subset of size 1,000 was randomly selected and used for model training, and the remaining 100 samples were used for prediction.

BSS was first applied to this example with the hyperparameter $\lambda = 2$ and three different choices of $n^* = 200$, 300 and 500. In simulations, we set $\sigma^2_{\theta_1} = 0.3$ and $\sigma^2_{\theta_2} = 0.5$, which have been calibrated such that the Markov chain can mix well in each run. The resulting parameter estimates and mean squared prediction errors (MSPE) for the prediction set were summarized in Table VII. The numerical results indicate that as n^* increases, BSS produces better prediction. It is also interesting to point out that as n^* increases, m tends to decrease when the same value of λ is used. This is reasonable, as the response variables can explain each other in the regression model we formulated. It is known that for the model (1.1), when the correlation function is exponential or Matérn, the parameters ϕ and σ^2 are non-estimable due to the existence of equivalent probability measures. However, in this case, the ratio ϕ/σ^2 is still estimable as shown in Zhang (2004). For this reason, we report in Table VII the estimate of the ratio σ^2/ϕ, instead of the respective estimates of σ^2 and ϕ. Our numerical results indicate that BSS produced accurate estimates of ϕ/σ^2 for this example. As a possible tool for determining n^*, we also reported in Table VII the mean squared fitting errors (MSFE_{t_1}) for the tier 1 neighboring observations.

Apparently, MSFE_{t_1} provides a good ordering for MSPE.

Table VII. Comparison of BSS and BFD method for the illustrative example. The number in the parenthesis denotes the standard error of the estimate. The CPU times were recorded for a single run of the algorithm on a desktop of Dual Core 3.0 GHz. BFD: Bayesian method for the full data; MSPE: mean squared prediction error; MSFE_{t_1}: mean squared fitting error for the tier 1 neighbors. Proportion was calculated in $(n^* + m)/n \times 100\%$.

	True	BSS(n^*, λ)			BFD
		$(200, 2)$	$(300, 2)$	$(500, 2)$	
m	—	37(0.21)	34(0.23)	28.9(0.19)	—
Proportion	—	23.7%	33.4%	52.9%	100%
ξ_0	0.5	0.54(0.09)	0.52(0.09)	0.56(0.09)	0.42(0.00)
ξ_1	1.0	0.97(0.01)	0.99(0.02)	1.00(0.00)	0.99(0.06)
ϕ/σ^2	25	26.58(2.22)	25.67(1.94)	24.83(1.38)	23.85(0.93)
τ^2	0.25	0.23(0.01)	0.24(0.01)	0.24 (0.01)	0.25(0.01)
MSPE	—	0.413(0.01)	0.398(0.01)	0.384(0.01)	0.381(0.01)
MSFE_{t_1}	—	0.449(0.01)	0.416(0.01)	0.395(0.01)	—
CPU(h)	—	0.5	1.5	7.3	47.8

For comparison, we also applied the standard Bayesian approach to this example. This approach works on the full dataset. Letting the parameters be subject to the priors (3.4) and (3.5), and integrating out $\boldsymbol{\xi}$ and σ^2, we get the posterior

$$f(\theta_1, \theta_2|D) \propto |R + \alpha I|^{-\frac{1}{2}} |\tilde{\Lambda}|^{-\frac{1}{2}} \frac{\Gamma(\frac{n}{2} + \epsilon)}{\{\boldsymbol{y}^T (R + \alpha I + \epsilon_\xi^{-2} CC^T)^{-1} \boldsymbol{y}/2 + \epsilon\}^{\frac{n}{2}+\epsilon}} \pi(\theta_1, \theta_2), \tag{3.12}$$

where R is the correlation matrix as defined in (1.2), $\tilde{\Lambda} = C^T (R + \alpha I)^{-1} C + \epsilon_\xi^2 I$, $\boldsymbol{y}$ is

an n-vector which consists of all observations in D, and

$$C = \begin{bmatrix} 1 & c_{s_1,1} & \cdots & c_{s_1,p} \\ \vdots & & \vdots & \\ 1 & c_{s_n,1} & \cdots & c_{s_n,p} \end{bmatrix},$$

is the design matrix of covariates. The Metropolis-with-Gibbs sampler is also applied to simulate from the posterior (3.12), but with only two parameters θ_1 and θ_2 updated at each iteration. The algorithm was also run once for each dataset. Each run consists of 10,000 iterations, where the first 5000 iterations were discarded for the burn-in process and 1000 samples were collected from the remaining iterations at equally-spaced time points. The resulting parameter estimates and the MSPE were reported in Table VII in the column of BFD (Bayesian method for Full Data). The simulation is very time consuming, as it needs to invert an $n \times n$ matrix at each iteration.

A comparison of the results from the two approaches indicates that although BSS costs much less CPU times than BFD, it can produce parameter estimates and prediction which both are as good as those produced by BFD. We note that the parameter estimates resultant from BSS may be biased due to the selection of $Y(\boldsymbol{s}^y)$ and inclusion of explanatory variables. For this example, this bias is ignorable because the prediction sites are randomly selected from the full dataset and the number of explanatory variables included in each model is relatively small. How to use BSS for parameter estimation will be discussed in the Chapter IV.

To understand why BSS works so well in both prediction and estimation, we conduct the following experiment to test if BSS can catch the long range dependence of the data. The experiment was done in the following procedure:

- For each sample in D_{-y} find its minimum distance to $\boldsymbol{s}^y$; that is, set

$$d(s) = \min_{s_i^y \in \boldsymbol{S}^y} \|s - s_i^y\|,$$

  for each site $s \in D_{-y}$.

- Divide the samples in D_{-y} into 10 groups according to the values of $d(s)$. Group 1 contains the one-tenth samples with the smallest values of $d(s)$, $\cdots$, and Group 10 contains one-tenth samples with the largest values of $d(s)$.

- Run BSS with $n^* = 500$ and $\lambda = 2$ for one dataset.

- Count the sampling frequency of the explanatory variables Z from each group.

Figure 6 shows the relative sampling frequency of the explanatory variables Z from each group. It indicates that, as expected, a high percentage of explanatory variables were drawn by BSS from the highly indexed groups, such as groups 8, 9 and 10. This indicates that BSS is indeed able to catch the long range dependence of the data. Therefore, it is understandable why BSS performs like BFD in estimation and prediction even with only a subset of the data being used.

To assess the sensitivity of BSS to the choice of λ, we tried different values of $\lambda = 1$, 2, 3, 5, and 10 for the case $n^* = 200$. The results were summarized in Table VIII. The results indicate that as λ increases, the number of explanatory variables included in the model tends to increase, the resulting regression model tends to be overfitted (the estimate of τ^2 tends to decrease slightly) and the contribution of covariates to the regression model tends to decrease (the estimate of ξ_1 tends to decrease). This experiment suggests that a small value of λ may be used, which will lead to a parsimony regression model in general.

In summary, the numerical results of this example suggests us to choose a reason-

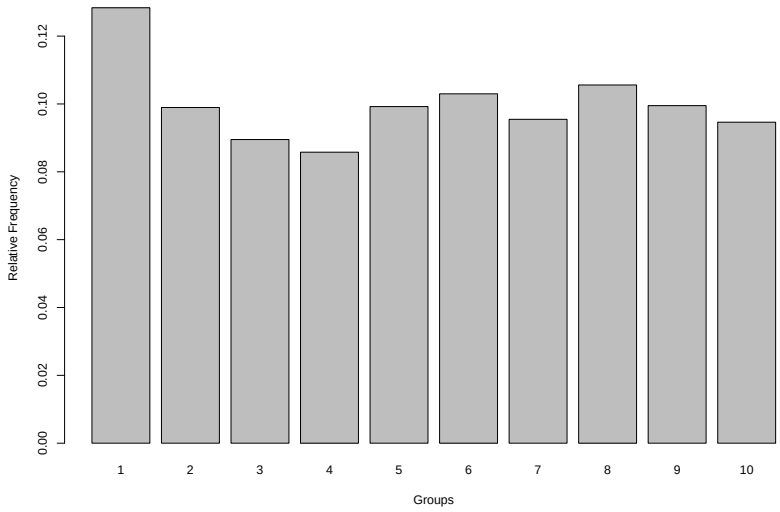

Fig. 6. Sampling frequency of the explanatory variables Z drawn by BSS for one dataset with $n^* = 500$ and $\lambda = 2$.

Table VIII. Sensitivity analysis for the value of λ. The number in the parenthesis denotes the standard error of the estimate. Proportion was calculated in $(n^* + m)/n \times 100\%$.

	Models (n^*, λ)				
	$(200, 1)$	$(200, 2)$	$(200, 3)$	$(200, 5)$	$(200, 10)$
m	26.5(0.17)	37(0.21)	45.8(0.27)	58(0.34)	82 (0.43)
Proportion	22.7%	23.7%	24.6%	25.8%	28.2%
ξ_0	0.52(0.08)	0.54(0.09)	0.53(0.09)	0.52(0.09)	0.50 (0.09)
ξ_1	0.98(0.01)	0.97(0.01)	0.96(0.01)	0.95(0.01)	0.94 (0.01)
ϕ/σ^2	26.06(2.16)	26.58(2.22)	25.11(2.39)	25.67(2.28)	24.52 (2.41)
τ^2	0.25(0.01)	0.23(0.01)	0.24(0.01)	0.23(0.01)	0.22 (0.10)
MSPE	0.414(0.01)	0.413(0.01)	0.414(0.01)	0.414(0.01)	0.415(0.01)

ably large value of n^* within the limit of our computer power, as a large value of n^* can generally work better in both parameter estimation and prediction. However, an excessively large value of n^* is not necessary, especially when one aims at prediction, as the prediction accuracy depends mainly on the neighbors of the prediction site. In practice, the value of n^* can be determined according to the value of MSFE_{t_1}. When n^* is reasonably large, say, the tier_3 neighboring points have been included in the response, a small value of λ, say, 1 or 2, may be used.

2. A Large Data Example

To assess the performance of BSS for large spatial data, we simulated 30 independent datasets from the model (1.1). Each dataset contains 20,100 observations with the sites uniformly distributed over the region $[0, 100] \times [0, 100]$. As for the last example, the data sets were generated using the function grf() in geoR (Ribeiro and Diggle, 2001) with the parameters $(\xi_0, \xi_1, \phi, \sigma^2, \tau^2) = (0.5, 1, 25, 1, 0.25)$ and the covariates generated from $N(0, 1)$. For each data set, 100 samples was randomly chosen and used for prediction, and the remaining 20,000 samples were used for model building.

BSS was applied to this example with $\sigma^2_{\theta_1} = \sigma^2_{\theta_2} = 0.3$, $\lambda = 1$, and $n^* = 300$, 500 and 700. The results were summarized in Table IX. The performance of BSS for this example is quite consistent with that for the last example. It produced very reasonable parameter estimates and MSPE values. For this example, we also calculated MSFE_{t_1}, the mean squared fitting errors for tier 1 neighboring observations. The results indicate again that MSFE_{t_1} is highly correlated with MSPE and can be used as a tool for choosing appropriate settings for BSS. It is worth pointing out that for this example, even with only less than 5% (on average) of samples being used at each iteration, BSS still performs reasonably well in both parameter estimation and prediction.

Table IX. Performance of BSS for the large data example. The estimates were calculated by averaging over the results from 30 different datasets and the number in the parentheses denotes the standard deviation of the estimate. Proportion was calculated in $(n^* + m)/n \times 100\%$.

	Models (n^*, λ)		
	$(300, 1)$	$(500, 1)$	$(700, 1)$
m	136(0.68)	134(0.98)	133(0.78)
Proportion	2.18%	3.17%	4.17%
ξ_0	0.665(0.100)	0.687(0.098)	0.705(0.096)
ξ_1	0.967(0.010)	0.985(0.006)	0.990(0.004)
ϕ/σ^2	23.05(1.86)	22.96(1.52)	23.39(1.58)
τ^2	0.228(0.007)	0.232(0.006)	0.237(0.004)
MSPE	0.345(0.00)	0.326(0.00)	0.316(0.00)
MSE_{t_1}	0.343(0.00)	0.320(0.00)	0.305(0.00)
Time(hr)	2.6	11.0	21.9

E. Real Data Study

1. Precipitation Anomaly Data

To demonstrate the performance of *BSS* for real problems, we considered a precipitation dataset from the National Climatic Data Center (NCDC) for the years 1895 to 1997. This data has been studied by many authors including Johns et al. (2003), Furrer et al. (2006), and Kaufman et al. (2008), among others. In this study, following Kaufman et al. (2008), we use the precipitation anomalies of 1962, available at `http://www.image.ucar.edu/Data/precip_tapering/`. This dataset consists of 7,352 samples (sites) and, as mentioned by Kaufman *et al.* (2008), there is no noticeable evidence for nonstationarity.

For this example, we randomly choose a subset of 250 out of 7,352 samples for model testing, and use the remaining samples for model building. We tried different values of $n^* = 250$, 500 and 750. Since our results reported in the previous section indicate that BSS is not sensitive to the value of λ, we set $\lambda = 1$ for this example. For each value of n^*, BSS was run 5 times independently with $\sigma^2_{\theta_1} = \sigma^2_{\theta_2} = 0.3$. Each run consisted of 10,000 iterations, with the first 5,000 iterations being discarded for the burn-in process and 1000 samples being collected from remaining 5,000 iterations at equally spaced time points. The results were summarized in Table X.

Table X shows an interesting pattern, the estimate of ϕ/σ^2 tends to decrease as n^* increases. This is reasonable. When $n^* = 250$, D_y consists of only the $Tier_1$ sites, which are far from each other. To establish the dependence among these sites, a large value of ϕ/σ^2 is needed. When n^* increases, the estimate of ϕ/σ^2 will converge to its true value. However, as long as n^* is reasonably large, say, $n^* \geq 3n^p$, BSS will perform very well in prediction. The reason is that the sparsity of neighboring information can be partially compensated by the updated parameter estimates. Table X shows

that BSS produced similar prediction results with $n^* = 500$ and $n^* = 750$ in terms of MSPE. Based on this observation, we conclude that BSS is a useful approach from a point of view of prediction.

To show that BSS can produce reasonable parameter estimates for model (1.1), we compare the predicted anomalies on a regular grid of 500×400 with the unit grid size (longitude $\times$ latitude) 0.065×0.12, where the anomalies were predicted using the covariance tapering method (Furrer *et al.*, 2006) with the BSS estimates given in Table X. In this study, we tapered the estimated covariance matrices by a spherical family with a range of 50 miles. The results were shown in Figure 7. The BSS prediction matches with observations very well, even for the case with $n^* = 250$. This indicates that the estimates produced by BSS are reasonable for this data. It needs to emphasize that BSS uses only a small proportion of the data at its each iteration.

Table X. BSS results for the anomalies of 1962. The estimates were calculated by averaging over the results of 5 independent runs, with their standard errors given in the parenthesis. The CPU times were recorded for a single run on a Desktop of Dual Core 3.0 GHz. Proportion was calculated in $(n^* + m)/n \times 100\%$.

	Models (n^*, λ)		
	$(250, 1)$	$(500, 1)$	$(750, 1)$
m	89(0.65)	90(0.59)	89(0.67)
Proportion	4.61%	8.03%	11.41%
ξ_0	-0.046(0.013)	-0.076(0.005)	-0.08(0.00)
ϕ/σ^2	206.16(11.10)	196.59(1.10)	172.76(1.93)
τ^2	0.096(0.011)	0.123(0.001)	0.112(0.001)
$MSPE$	0.320(0.003)	0.272(0.001)	0.272(0.000)
Time(hr)	1.3	7.8	24.8

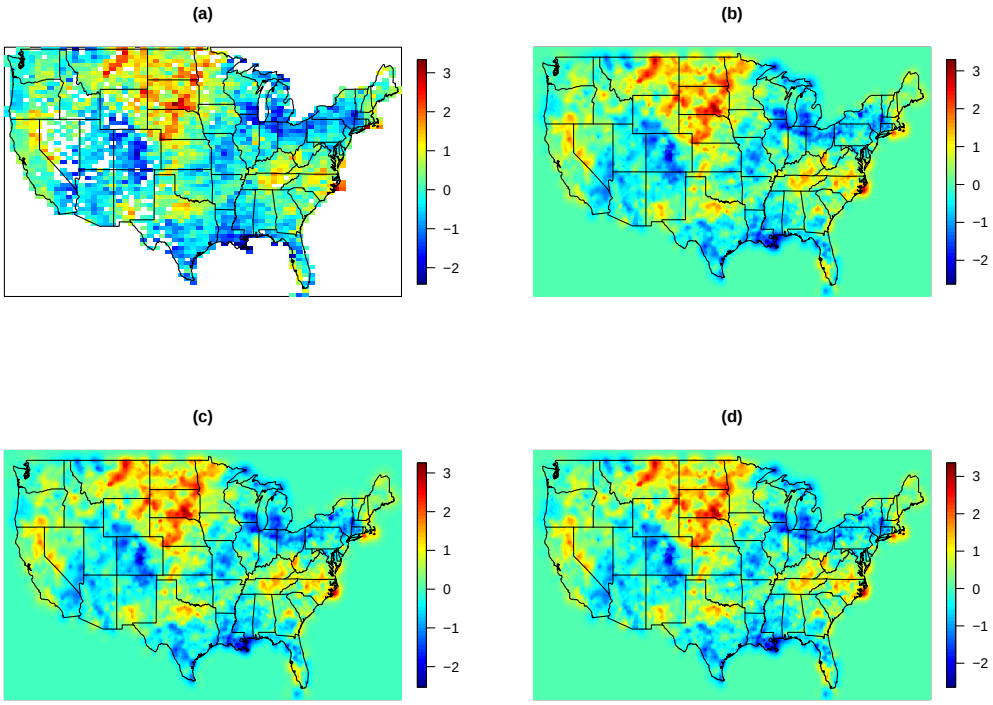

Fig. 7. Images of observed and predicted anomalies of 1962 on a regular grid of size 500×400. (a) Observed anomalies; (b) prediction surface for $n^* = 250$; (c) Prediction surface for $n^* = 500$; (d) prediction surface for $n^* = 750$.

2. Gold Mine Data

The Gold mine data, available at `http://www.kriging.com/datasets/`, is constructed based on a Wits type gold mine. The samples are chipped from the face of the reef in a working section of the mine (stope). As the face advances, new chip samples are taken. Values within a stope are traditionally estimated using the sample values from the face. The data set was used in Clark and Harper (2000). To ensure the data normality holds for model (1.1), we work on the logarithm of the observations.

The data set consists of 21,577 observations. We randomly select 250 observations

for model testing and use the remaining observations for model building. BSS was run for 5 times independently with $\sigma^2_{\theta_1} = 0.2$ and $\sigma^2_{\theta_2} = 0.3$. Each run consists of 10,000 iterations, where the first 5,000 iterations were discarded for the burn-in and 1,000 samples were collected from the remaining iterations at equally-spaced time points. The numerical results were summarized in Table XI.

Table XI shows a similar pattern to Table X: As n^* increases, the estimate of ϕ/σ^2 tends to decrease. Figure 8 shows the images of the observations and prediction surfaces. It indicates again that BSS can produce reasonable parameter estimates for model (1.1), even with only a small proportion (less than 5%) of the data being used at each iteration.

Table XI. BSS results for the gold mine data. The estimates were calculated by averaging over the results of 5 independent runs, with their standard errors given in the parenthesis. The CPU times were recorded for a single run on a Desktop of Dual Core 3.0 GHz. Proportion was calculated in $(n^* + m)/n \times 100\%$.

	Models (n^*, λ)	
	$(500, 1)$	$(750, 1)$
m	151(0.76)	152(1.38)
Proportion	3.02%	4.18%
ξ_0	3.76(0.003)	3.77(0.002)
$\phi/\sigma 2$	99.45(1.08)	71.21(1.19)
$\tau 2$	0.098(0.001)	0.058(0.001)
$MSPE$	0.154(0.000)	0.139(0.000)
Time(hr)	9.4	28.0

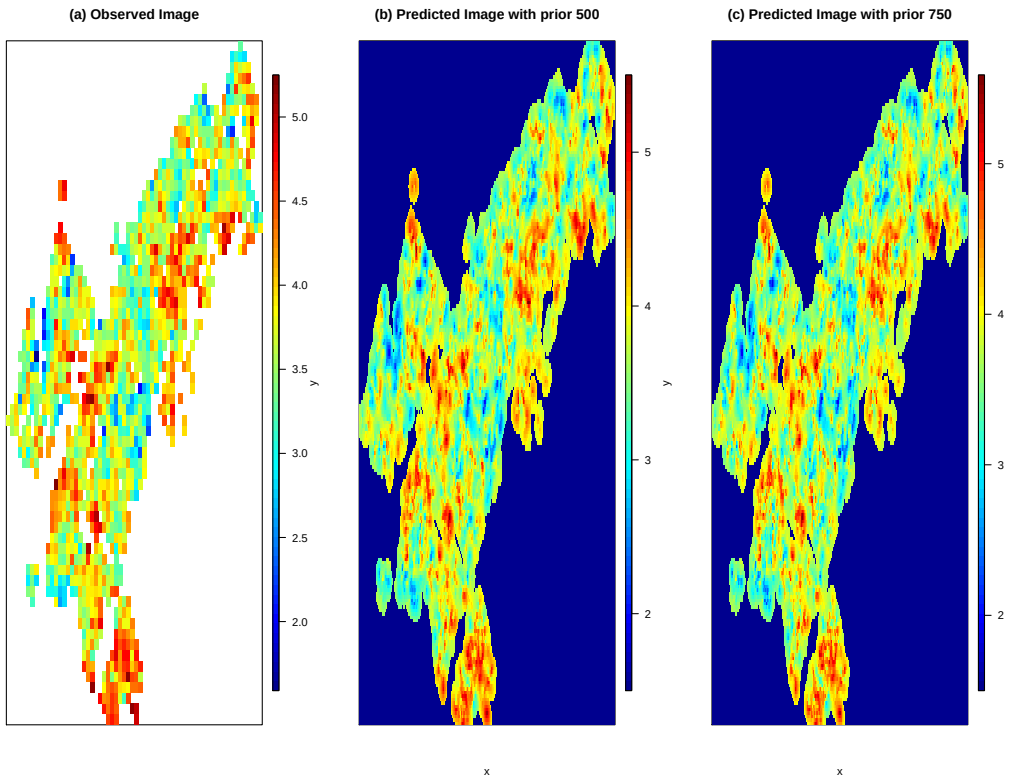

Fig. 8. Images of observations and predicted surfaces on a regular grid of size 300×200 for the goldmine data. The prediction surfaces were produced by local Kriging for which each grid point is predicted based on the nearest 100 points. (a) Images of observations; (b) prediction surface by the BSS estimate with $n^* = 500$; and (c) prediction surface by the BSS estimate with $n^* = 750$.

CHAPTER IV

SUMMARY AND DISCUSSION

In this dissertation, we propose two approaches to address computational issues of Gaussian geostatistical models, namely, the auxiliary lattice model (ALM) approach and the Bayesian site selection (BSS) approach. The key feature of ALM is to introduce a latent regular lattice which links Gaussian Markov Random Field (GMRF) with Gaussian Field (GF) of the observations. The GMRF on the auxiliary lattice represents an approximation to the Gaussian process $\{X(s)\}$. It is remarkable that the computational complexity of ALM is only $O(n)$, which implies that our model can be applied to very large data sets with reasonable CPU times. The numerical results show that ALM can approximate Gaussian random fields very well, even for those with long correlation lengths. For real data examples, ALM can generally outperform conventional Gaussian random field models in both prediction errors and CPU times.

ALM approach is thought to work efficiently under the situation that a density of the observation site is uniform over the region of interest (ROI) because the optimal lattice size depends on the density of observation sites. In the case that a distribution of observation sites vary over the ROI, the optimal lattice size should accordingly vary so that the ALM approach based on a uniform lattice size may make it less efficient.

ALM can be extended in various ways. Here are some examples;

- Under the framework of our model, anisotropy can be easily incorporated into our model. For example, if we allow the interaction parameters β_v, β_h and β_d to take different values, the resulting model will be of anisotropy.

- Under the framework of our model, nonstationarity can also be easily incorporated into our model. Instead of assuming that the hidden Gaussian process

$\{X(s_i)\}$ has a constant variance over all sites, we can model $\log\{\sigma^2(s_i)\}$, the log-variance of $X(s_i)$, as a spline function of the sites. For example, we may set

$$\log\{\sigma^2(s_i)\} = \gamma_1 B_0(s_{1i}, s_{2i}) + \cdots + \gamma_k B_K(s_{1i}, s_{2i})$$

where $s_i = (s_{1i}, s_{2i})$ denotes the coordinate of the site s_i, k denotes the number of knots, $B(\cdot)$ denotes the B-spline basis function, and $(\gamma_1, \cdots, \gamma_k)$ is the vector of coefficients.

- ALM can be extended by allowing the auxiliary GMRF Z to have a higher-order neighborhood structure. As discussed in Section A of Chapter II, this extension is straightforward. In addition, our model can be easily applied to spatial-temporal data by introducing a three-dimensional auxiliary lattice to the time-space data. In this case, the eigenvalues of the auxiliary GMRF are also available analytically using the technique of discrete Fourier transformations. (See e.g., Allcroft and Glasbey, 2003 for the details).

- In this dissertation, our model is illustrated with only the spherical correlation function. The spherical correlation function is computationally efficient for the model M_{AL} due to its bounded support. However, a compact support of the correlation function is not an essential requirement for our model. We note here that for the model M_{FAL} all types of correlation functions are equally efficient in computation, as the model employs a fixed neighborhood system.

- Regarding computation of our model, we note that the method of integrated nested Laplace approximation (INLA) (Rue *et al.*, 2009) can also be applied for posterior inference as an alternative way to MCMC simulations. However, when the model is extended to include more parameters for handling other features of the data, such as isotropy, nonstationarity or higher-order neighboring

dependence, the applicability of INLA may be questionable, as INLA is best suited to the models for which the number of parameters is low.

- In this dissertation, we suggest to choose the size of auxiliary lattice through a cross-validation approach. A more Bayesian approach, namely, the DIC approach (Speigelhalter *et al.*, 2002), can also be applied to determine the size of auxiliary lattice. With the DIC approach, the split of training and test sets can be avoided for the available data.

In addition to methodological extensions of our model, a further theoretical study on the feature of our model is also of great interest. For example, how does the auxiliary lattice affect the parameter estimation of our model, and does the auxiliary lattice model suffer from a nonidentifiability problem as its parent model (1.1) for certain correlation functions ?

The second approach, Bayesian site selection (BSS), attempts to reduce the dimension of data through a smart selection of a representative subset of the observations while keeping data information not lost significantly. The BSS approach works by performing a regression analysis based on the prediction request, with the data dimension being reduced through a stochastic variable selection procedure. Our simulated examples show that with an appropriate choice of response variables and an appropriate choice of λ, BSS can produce parameter estimates and prediction which both are nearly as good as those produced by the Bayesian method with the full data, although BSS uses only a small proportion of the data at each iteration. For a really large data set, say, the number of observations is over 20,000, our numerical results (of the 2nd simulated example and the 2nd real example) indicate that BSS can produce very reasonable parameter estimates and predictions with only less than 5% of the data used at each iteration.

BSS approach is thought to work efficiently, as is with ALM, under the situation where a density of the observation sites are uniformly distributed over ROI. BSS may have less prediction ability like other algorithms for the site having a sparser observation neighborhood because the neighboring observations play a major role in prediction.

As previously mentioned, the parameter estimates produced by BSS can be biased due to the choice of the response variables and inclusion of explanatory variables. For example, when the response variables are not uniformly selected from the set of observations and the number of explanatory variables included in the regression is too large, the resulting parameter estimates may be biased. To address this issue, we propose the an ensemble BSS approach, which works in a style of bootstrap sampling as follows:

- Select multiple response sets, with each being drawn randomly from the set of observations.
- Run BSS for each response set.
- Average the parameter estimates resultant from each response set.

In this case, the hyperparameter λ may be set to a small number or even zero, as one aims at parameter estimation instead of prediction. Following from the standard theory of bootstrap, the parameter estimates resultant from the ensemble BSS approach is unbiased.

REFERENCES

Allcroft, D.J. and Glasbey, C.A. (2003), "A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation," *Applied Statistics*, 52, 487-498.

Balram, N. and Moura, J.M.F. (1993), "Noncausal Gauss Markov random fields: Parameter structure and estimation," *IEEE Transactions on Information Theory*, 39, 1333-1355.

Banerjee, S., Gelfand, A.E., Finley, A.O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society*, Ser. B, 70, 825-848.

Besag, J. and Modal, D. (2005), "First-order intrinsic autoregressions and the Wijs process," *Biometrika*, 92, 909-920.

Billings, S.D., Newsam, G.N. and Beatson, R.K. (2002), "Gaussian predictive process models for large spatial data sets," *Geophysics*, 67, 1823-1834.

Clark, I. and Harper, V.W. (2000), *Practical Geostatistics*, Ecosse North America Llc, p. 4.

Cressie, N. (1993), *Statistics for Spatial Data* (Second Edition), New York: John Wiley. pp. 61-62, 85-56.

Diggle, P.J., Tawn, J.A., and Moyeed, R.A. (1998), "Model based geostatistics "(with discussion), *Applied Statistics*, 47, 299-350.

Finely, A.O., Sang H., Banerjee, S. and Gelfand, A.E. (2009), "Improving the performance of predictive process modeling for large datasets," *Computational Statistics & Data Analysis*, 53, 2873-2884.

Fuentes, M. (2007), "Approximate likelihood for large irregularly spaced spatial data," *Journal of the American Statistical Association*, 102, 321-331.

Furrer, R., Genton, M. G. and Nychka, D. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15, 502-523.

Furrer, R. and Bengtsson, T. (2007), "Estimation of High-dimensional Prior and Posterior Covariance Matrices in Kalman Filter Variants," *Journal of Multivariate Analysis*, 98, 227-255.

Hartman, L. and Hössjer, O. (2008), "Fast kriging of large data sets with Gaussian Markov random fields," *Computational Statistics and Data Analysis*, 52, 2331-2349.

Jones, C.J., Nychka, D., Kittel, T.G.T. and Daly, C. (2003), "Infilling Sparse Records of Spatial Fields," *Journal of the American Statistical Association*, 98, 796-806.

Jones, R.H. and Zhang, Y. (1997), "Models for continuous stationary space-time processes," in *Modeling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions*, eds. P.J. Diggle, W.G. Warren and R.D. Wolfinger, New York: Springer-Verlag.

Kammann, E.E. and Wand, M.P. (2003), "Geoadditive models," *Applied Statistics*, 52, 1-18.

Kaufman, C., Schervish, M., and Nychka, D. (2008), "Covariance Tapering for Likelihood-Based Estimation in Large Spatial Datasets," *Journal of the American Statistical Association*, 103, 1156-1569.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (2000), "Smoothing spline ANOVA models for large datasets with Bernoulli observations and the randomized GACV," *Annals of Statistics*, 28, 1570-1600.

Lindgren, F., Rue, H. and Lindström, J. (2011), "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach," *Journal of the Royal Statistical Society*, Ser. B, 73, 423498.

Müller, P. (1991), "A generic approach to posterior integration and Gibbs sampling," Technical report, Purdue University, West Lafayette, Indiana.

Paciorek, C.J. (2007), "Computational techniques for spatial logistic regression with large datasets," *Computnl Statist. Data Anal.*, 51, 3631-3653.

Ribeiro Jr., P.J. and Diggle, P.J. (2001) *geoR: A package for geostatistical analysis*, R-NEWS Vol1, No 2, ISSN 1609-3631.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC.

Rue, H., Martino, S. and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society*, Ser. B, 71, 319-392.

Rue, H. and Tjelmeland, H. (2002), "Fitting Gaussian Markov random fields to Gaussian field," *Scandinavian Journal of Statistics*, 29, 31-49.

Searle, S.R. (1982), *Matrix Algebra Useful for Statistics*, John Wiley & Sons, Inc., p. 313.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002), "Bayesian measures of model complexity and fit"(with discussion), *Journal of the Royal Statistical Society*, Ser. B, 64, 583-639.

Stein, M.L. Chi, Z. and Welty, L.J. (2004), "Approximating likelihoods for large spatial data sets," *Journal of the Royal Statistical Society*, Ser. B, 66, 275-296.

Sun, Y., Li, B. and Genton, M. G. (2012), "Geostatistics for Large Datasets", in *Space-Time Processes and Challenges Related to Environmental Problems*, eds. E. Porcu, J. M. Montero and M. Schlather, Springer, Vol. 207, Chapter 3, pp. 55-77, available at *http://www.stat.tamu.edu/ genton/2012.SLG.chapter3.final.pdf.*

Tanner, M. and Wong, W.H. (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 559-568.

Vecchia, A.V. (1988), "Estimation and model identification for continuous spatial processes," *Journal of the Royal Statistical Society*, Ser. B, 50, 297-312.

Wikle, C. and Cressie, N. (1999), "A dimension-reduced approach to space-time Kalman filtering," *Biometrika*, 86, 815-829.

Zhang, H. (2004), "Inconsistent estimation and asymptotically equal interpolations in model-baed geostatistics," *Journal of the American Statistical Association*, 99, 250-261.

APPENDIX A

PROOF OF THEOREM 1

Let

$$B = \begin{bmatrix} Q^{-1} & \boldsymbol{r} \\ \boldsymbol{r}^T & 1 \end{bmatrix},$$

be an $n \times n$ symmetric matrix, where Q is positive definite. To show the theorem, it suffices to show that B is positive definite if $1 - \boldsymbol{r}^T Q \boldsymbol{r} > 0$. For the proof, we need the following lemma.

Lemma 2 *(Searle, 1982, Theorem 2) For symmetric matrices A and B of the same order, with A being positive definite, there exists a nonsingular matrix P such that $P^T A P = I$ and $P^T B P$ is a diagonal matrix of the solutions for λ to $\|B - \lambda A\| = 0$.*

In what follows, we consider two cases: (i) $n = 2$ and (ii) $n > 2$.

(i) If $n = 2$, then Q^{-1} is a scalar. In this case, we have

$$|B - \lambda I| = \lambda^2 - \left(\frac{1}{Q} + 1\right)\lambda + \frac{1}{Q} - \boldsymbol{r}^2.$$

Let λ_1 and λ_2 denote the two roots of $|B - \lambda I| = 0$. The condition $1 - \boldsymbol{r}' Q \boldsymbol{r} > 0$ implies $\frac{1}{Q} - \boldsymbol{r}^2 > 0$; that is, $\lambda_1 \lambda_2 > 0$. In addition, we have $\lambda_1 + \lambda_2 = \frac{1}{Q} + 1 > 0$. Hence, we must have $\lambda_1 > 0$ and $\lambda_2 > 0$; that is, B is positive definite.

(ii) If $n > 2$, then Q is a matrix. In this case, we define

$$A = \begin{bmatrix} Q^{-1} & 0 \\ 0 & 1 \end{bmatrix}, A^* = \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & 1 \end{bmatrix},$$

which are both positive definite. By Lemma 2, there exists a nonsingular matrix P such that $P^TAP = I$ and $P^TBP = Diag[\lambda_1, \ldots, \lambda_n]$, where λ_i's are the solutions to $|B-\lambda A| = 0$. Since $|B-\lambda A| = 0$ is equivalent to $|A^*||B-\lambda A||A^*| = 0$, we consider the matrix

$$C = A^*(B - \lambda A)A^* = \begin{pmatrix} (1-\lambda)I & Q^{\frac{1}{2}}\boldsymbol{r} \\ \boldsymbol{r}^TQ^{\frac{1}{2}} & 1-\lambda \end{pmatrix}$$

If $\lambda = 1$, then $\text{rank}(C) \leq 2$, so that $|C| = 0$. Hence, $\lambda = 1$ is a solution to $|B - \lambda A| = 0$.

If $\lambda \neq 1$, then

$$|C| = (1-\lambda)^{n-2}[(1-\lambda)^2 - \boldsymbol{r}^TQ\boldsymbol{r}].$$

So the solutions to $|B-\lambda A| = 0$ are $1 \pm \sqrt{\boldsymbol{r}^TQ\boldsymbol{r}}$. Therefore, all possible solutions to $|B - \lambda A| = 0$ are only 1, $1 + \sqrt{\boldsymbol{r}^TQ\boldsymbol{r}}$, and $1 - \sqrt{\boldsymbol{r}^TQ\boldsymbol{r}}$. This implies that B is positive definite.

APPENDIX B

DERIVATION OF (2.17)

Consider the set-up in Section C of Chapter II. For $1 \leq i \leq n$, we can rewrite

$$\mu_i(z_{kl}^*) \equiv \mu_i(z_{kl}^*, \boldsymbol{z}_{-(kl)}) = (\boldsymbol{r}_i^T Q(\boldsymbol{\beta}))_{kl} z_{kl}^* + \alpha_{kl}(i),$$

where $(\boldsymbol{r}_i^T Q(\boldsymbol{\beta}))_{kl}$ denotes the $((k-1) \times N + l)^{th}$ element of the vector $\boldsymbol{r}_i^T Q(\boldsymbol{\beta})$, and

$$\alpha_{kl}(i) = \sum_{ab \neq kl} (\boldsymbol{r}_i^T Q(\boldsymbol{\beta}))_{ab} z_{ab}.$$

Then it follows that

$$\begin{aligned} f(z_{kl}^* | \boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}, \boldsymbol{y}) &\propto f(z_{kl}^*, \boldsymbol{z}_{-(kl)}, \boldsymbol{\theta} | \boldsymbol{y}) \\ &\propto f(z_{kl}^*, \boldsymbol{z}_{-(kl)} | \sigma^2, \boldsymbol{\beta}) f(\boldsymbol{y} | z_{kl}^*, \boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}) \\ &\propto f(z_{kl}^* | \boldsymbol{z}_{-(kl)}, \sigma^2, \boldsymbol{\beta}) f(\boldsymbol{y} | z_{kl}^*, \boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}). \end{aligned}$$

The density $f(z_{kl}^* | z_{-(kl)}, \sigma^2, \boldsymbol{\beta})$ can be obtained in Rue and Tjelmeland(2002) by

$$Z_{kl}^* | \boldsymbol{z}_{-(kl)}, \sigma^2, \boldsymbol{\beta} \sim N \left(\beta_h \sum_{(k',l') \in \partial_{h(kl)}} z_{k'l'} + \beta_v \sum_{(k',l') \in \partial_{v(kl)}} z_{k'l'} + \beta_d \sum_{(k',l') \in \partial_{d(kl)}} z_{k'l'}, \sigma^2 \right),$$

and $f(\boldsymbol{y} | z_{kl}^*, \boldsymbol{z}_{-(kl)}, \boldsymbol{\theta})$ is given in (2.8).

Now that

$$\begin{aligned} f(\boldsymbol{y} | z_{kl}^*, \boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}) &\propto \prod_{i=1}^n \exp \left[-\frac{1}{2(\tau^2 + \sigma_i^2)} \{ y(s_i) - \nu(s_i) - \mu_i \}^2 \right] \\ &\propto \prod_{i=1}^n \exp \left[-\frac{1}{2(\tau^2 + \sigma_i^2)} \left\{ y(s_i) - \nu(s_i) - \alpha_{kl}(i) - \boldsymbol{r}_i^T Q(\boldsymbol{\beta})_{kl} z_{kl}^* \right\}^2 \right] \\ &\propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \frac{(\boldsymbol{r}_i^T Q(\boldsymbol{\beta})_{kl})^2}{\tau^2 + \sigma_i^2} z_{kl}^{*2} + \sum_{i=1}^n \frac{y(s_i) - \nu(s_i) - \alpha_{kl}(i)}{\tau^2 + \sigma_i^2} \boldsymbol{r}_i^T Q(\boldsymbol{\beta})_{kl} z_{kl}^* \right\}, \end{aligned}$$

it follows that

$$f(z^*_{kl}|\boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2E^2} z^{*2}_{ij} + F z^*_{ij}\right\},$$

so that $Z^*_{kl}|\boldsymbol{z}_{-(kl)}, \boldsymbol{\theta}, \boldsymbol{y} \sim N(E^2F, E^2)$ where E^2 and F are defined in (2.18).

APPENDIX C

SIMULATION STUDY FOR LONG LENGTH CORRELATION CASE

To illustrate the performance of our models for the data with long range correlations, we simulated 10 data sets of size 1,500 with $\phi = 40$ and other parameters being set as in Section E.2 of Chapter II. As shown in Section E.2, M_{FAL} can performs as well as M_{AL} in prediction, so we consider only M_{FAL} in this subsection. For the model M_{FAL}, we tried four different lattices, 20×20, 30×30, 40×40 and 50×50. For each dataset, the algorithm was run for 21,000 iterations, for which the first 1,000 iterations were discarded for the burn-in process and then 1,000 samples were collected from the remaining iterations at equally-spaced time points. For comparison, the model M_G was also applied to this example with the same prior setting as specified in the previous section. The numerical results, reported in Table XII, indicate that M_{FAL} can approximate the Gaussian Markov random field very well in terms of predictions even when the correlation length is long. Note that all comments made for the case of short length correlation are also applicable here.

Table XII. Parameter estimation of the models M_{FAL}, and M_G for the simulated data with a long correlation length of $\phi = 40$. The number in the parentheses denotes the standard deviation of the estimate. CPU: Measured in minutes on a 3.0 GHz Intel Core 2 Duo computer for a single run for one dataset.

	M_{FAL}				M_G
	20×20	30×30	40×40	50×50	
β	0.121(0.001)	0.124(0.000)	0.124(0.000)	0.124(0.000)	
ϕ	15.19(0.41)	10.04(0.31)	7.49(0.06)	6.01(0.03)	50.31(13.44)
ν	1.00(0.96)	1.01(1.01)	1.02(1.02)	1.02(1.02)	0.89(0.68)
σ^2	3.67(0.43)	2.33(0.24)	1.79(0.20)	1.57(0.16)	8.70(2.39)
τ^2	0.84(0.16)	0.94(0.16)	0.96(0.10)	0.90(0.11)	1.13 (0.19)
MSPE	1.77(0.09)	1.71(0.10)	1.72(0.11)	1.72(0.11)	1.68(0.09)
CPU(m)	2.26	2.80	2.51	2.83	101.7

VITA

- Name

  Jincheol Park

- Address

  Department of Statistics, Texas A & M University

  3143 TAMU, College Station, TX 77843-3143

- Email

  jpark@stat.tamu.edu

- Education

  MSc., Statistics, University of Ottawa, 2000, Canada

  BSc., Statistics, Seoul National University, 1997, South Korea